VERİ BİLİMİ DERGİSİ
www.dergipark.gov.tr/veri

Ortaokul Öğrencilerinin Akademik Başarılarının Eğitsel Veri Madenciliği Yöntemleri ile Tahmini

Banu ABBASOĞLU[1]
Bahçeşehir Üniversitesi, Eğitim Bilimleri Enstitüsü, Bilgisayar ve Öğretim Teknolojileri, İstanbul

Makale Bilgisi

Başvuru
12/06/2020
Kabul:
11/07/2020

Özet

Eğitsel Veri Madenciliği, eğitim ortamlarından gelen benzersiz veri türlerini araştırmak için yöntemler geliştirmek, öğrencileri ve öğrendikleri ortamları daha iyi anlamak için bu yöntemleri kullanmakla ilgilenen yeni bir disiplindir. Eğitsel veri madenciliği, bilgisayar bilimi, eğitim ve istatistik alanlarının birleşimi olarak düşünülebilir. Bu çalışmanın amacı, öğrencilerin demografik özelliklerinin ve sosyoekonomik durumlarının öğrencilerin yıl sonu genel başarı ortalamalarına olan etkilerini eğitsel veri madenciliği yöntemleri ile analiz etmektir. Bu amaçla, 2019-2020 eğitim-öğretim yılı, 2. Dönemi'nde, Yalova ilinde sosyo demografik açıdan farklı dört resmi ortaokuldaki, 5, 6, 7 ve 8. sınıf, 1395 ortaokul öğrencisinin, E-Okul Yönetim Bilgi Sisteminden sosyo demografik özelliklerine ilişkin verileri elde edilmiştir. Daha sonra elde edilen verilerden sınıflandırma teknikleri ve algoritmaları ile yıl sonu genel başarım ortalamaları tahmin edilmiştir. Sınıflandırıcı algoritmaların uygulanması sonucunda yıl sonu genel başarı ortalaması başarımında *lojistik* algoritması en iyi tahmini gerçekleştirmiştir.

Anahtar Kelimeler: *Eğitsel Veri Madenciliği, Ders Başarı Ortalaması Tahmini, Sınıflandırma*

Prediction of Academic Achievements of Secondary School Students with Educational Data Mining Methods

Abstract

Educational Data Mining is a new discipline that is interested in developing methods to explore unique data types from educational environments, and using these methods to better understand students and the environments they learn. It can be thought of as a combination of educational data mining, computer science, education and statistics. The aim of this study is to analyze the effects of demographic characteristics and socioeconomic status of students on the overall average success scores of students by using educational data mining methods. For this purpose, in the second term of 2019-2020 academic year, data on sociodemographic properties of 1395 middle school students from class 5, 6, 7 and 8 in Yalova province via e-School Management Information System were obtained. Afterwards, the average of year-end overall performance with classification techniques and algorithms was estimated. As a result of the application of classifier algorithms, the logistic algorithm has achieved the best estimation in the performance of the end-of-year overall success average.

Keywords: *Educational Data Mining, Course Success Average Estimation, Classification*

* iletişim e-posta: banu.abbasoglu@gmail.com
** Bu çalışmanın bir kısmı III. International Conference on Data Science and Applications 2020'de sözlü olarak sunulmuştur.

1 Giriş

Bourdieu'nun ünlü kültürel yeniden üretim teorisindeki temel hipotez, nesiller boyunca aktarılan, aileler ve bireyler tarafından sahip olunan kültürel sermayenin, bireylerin eğitim başarısına katkıda bulunan önemli bir kaynak olduğudur, yüksek sosyoekonomik düzeye sahip aileler çocuklarına daha fazla eğitim kaynağı sağlar ve ergenlerin eğitimsel başarısını teşvik eder [1, 2]. Sosyoekonomik düzeyi yüksek olan bir aile, çocukları için daha iyi bir yaşam ortamı ve daha fazla eğitim kaynağı sağlayabilir [3]. Sosyoloji ve eğitim alanında, ailenin sosyoekonomik durumu ile akademik başarı arasındaki ilişkiyi belirlemeye yönelik pek çok araştırma yapılmıştır. Literatürün kapsamlı incelenmesi sonucunda sosyoekonomik durum ile akademik başarı arasında anlamlı bir ilişkinin olduğu birçok araştırma sonuçlarına rastlanmıştır. Örneğin; ABD'deki dört ilköğretim okuluna kayıtlı 8.sınıf öğrencilerinin sosyoekonomik düzeyinin akademik başarılarına olan etkisi araştırılmıştır. Bu araştırmada, öğrenciler ekonomik olarak dezavantajlı ve dezavantajlı olmayan öğrenciler olarak kategorize edilmiştir. Sonuçlar ekonomik olarak dezavantajlı öğrencilerin matematik, dil sanatları, sosyal bilgiler ve fen puanlarının ekonomik olarak dezavantajlı olmayan öğrencilere göre daha düşük olduğunu göstermiştir [4]. Yine Pakistan'daki 1580 ortaokul öğrencisinin sosyoekonomik düzeyinin akademik başarıları üzerindeki etkileri araştırılmıştır. Araştırmacı öğrencileri üst, orta ve düşük sosyal sınıflara ayırmış, bulgular, üst sınıfa mensup öğrencilerin orta ve düşük sınıfa mensup öğrencilerinden daha başarılı olduğunu göstermiştir [5]. Türkiye'de ise yapılan bir çalışmada, 8. sınıf ilköğretim okulu öğrencilerine, 25 soruluk bir anket uygulanmış ve çeşitli değişkenlerin (annelerin eğitim düzeyi, kardeş sayısı) akademik başarı üzerindeki etkisini araştırılmıştır. Araştırmacı, öğrencilerin evdeki olanakları ve annelerin eğitim düzeylerinin artması ve kardeşlerinin sayısının azalmasıyla öğrencilerin okullarda akademik performansta artış gösterdiğini tespit etmiştir [6]. Bir başka çalışmada, 1990 ve 2000 yılları arasında yayınlanan dergilerdeki sosyoekonomik durum ve akademik başarı ile ilgili literatürü gözden geçirilmiş, incelenen sonuçlar ışığında ebeveynlerin sosyoekonomik yapıdaki konumlarının öğrencilerin akademik başarıları üzerinde güçlü bir etkisi olduğunu göstermiştir [7]. Bu bağlamda akademik başarı ile sosyoekonomik durumun arka plan ilişkisi ile ilgili birçok çalışma yapıldığı ve ilişkinin güçlü yönde olduğu söylenebilir. Öğrencilerin akademik başarılarında pek çok faktör rol oynar. Şimdiye kadar yapılan çalışmalardan öğrencilerin akademik başarısında hangi faktörlerin öncelikli olarak rol oynadığı tam olarak ortaya konulamamıştır. Ülkemizde eğitime yapılan teknolojik, fiziki ve proje bazlı yatırımlara rağmen uluslararası sınavlarda başarı istenen seviyede değildir. Bireyin akademik başarısı üzerinde etkili olan pek çok değişken bulunabilir. Bu araştırmada bireyin akademik başarısı üzerinde etkisi olabileceği düşünülen sosyoekonomik (anne baba eğitimi, anne ve babanın hayatta olma durumu, aile ile yaşama durumu, ailenin ekonomik geliri, ailedeki kardeş sayısı, odası olma durumu, takviye kurs alma durumu) ve demografik (yaş, cinsiyet, sürekli hastalık durumu, özel eğitim durumu, devam durumu) verileri eğitsel veri madenciliği yöntemleri ile işlenerek bireyin akademik başarım analizinin yapılması amaçlanmıştır.

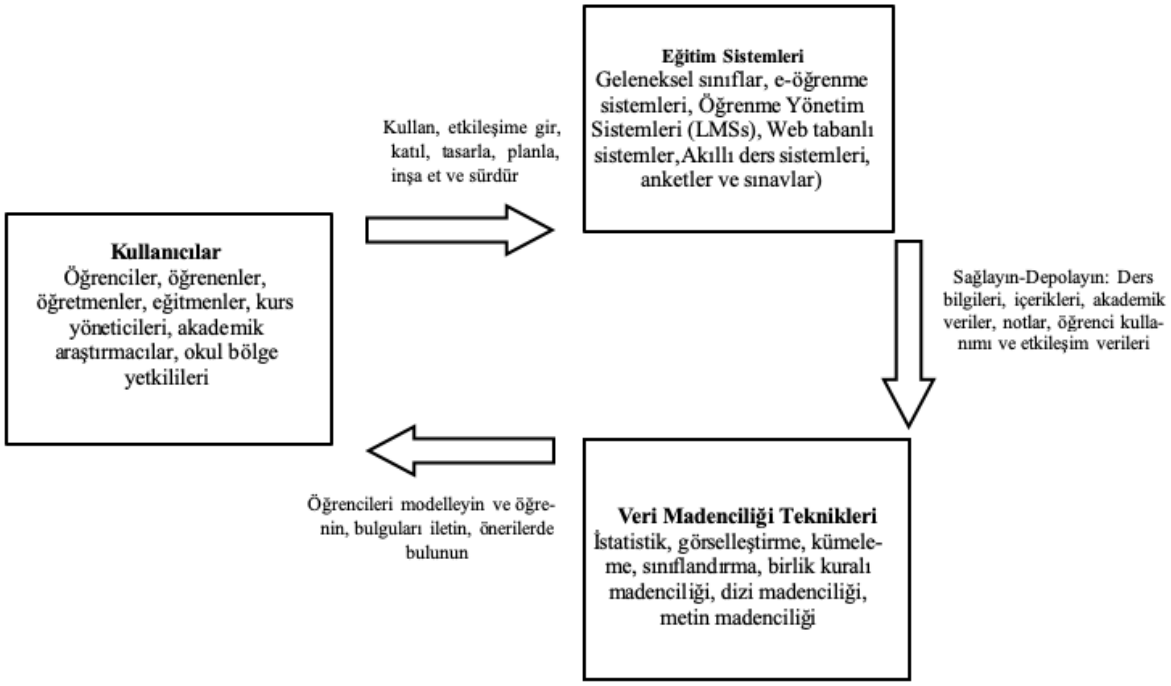
Eğitsel veri madenciliği, geleneksel, açık ve uzaktan eğitim ortamlarındaki verileri araştırmak için tasarım modelleri, yöntemler ve algoritmalara yönelik bir paradigma olarak ortaya çıkmıştır. Meksika'da eğitimin kalitesini artırmak için, geleneksel, açık ve uzaktan eğitim ortamlarından elde edilen verilerle, sınıflandırma (Bayes teoremi, karar ağaçları, J48) kümeleme ve regresyon yöntemleri kullanılarak öğrenci davranış ve performanslarının modellemesi, değerlendirilmesi ve bu değerlendirmeler üzerinden geribildirim sağlanması, müfredat ve öğretmenlerin değerlendirilmesi, bilgi keşfi amacıyla, daha doğrusu yeni bir eğitim reformu yaratmak için eğitsel veri madenciliğinden yararlanılmıştır [8].

Ülkemizde yapılan bir çalışmada, 6, 7, 8.sınıf ortaokul öğrencilerine 24 soruluk bir anket uygulamış, Türkçe, Matematik dersleri ve dönem sonu genel başarı ortalamalarını regresyon / çok sınıflı makine öğrenmesi modelleri oluşturarak tahmin etmiştir [9]. Uzaktan eğitimde de benzer amaçlar doğrultusunda çalışmalar gerçekleştirilmiştir. Transilvanya Üniversitesi'nde gerçekleştirilen çalışmada bilgisayar kullanım alanı, bilgisayar kullanımının önemi, bilgisayar kullanımını gerektiren fakülte etkinlikleri, öğrencilerin üniversitede bilgisayar kullanım süreleri, internet ve web sitesi kullanımı, teknoloji yatırımları, BT kaynaklarına erişim konularıyla ilgili öğrencilerin eğitim sisteminin bilgisayarlaşmasıyla ilgili görüşlerini analiz etmiştir [10]. Eğitsel veri madenciliğinde ulusal ve uluslararası farklı öz nitelikler kullanılarak yapılan çalışmalar olduğu görülmektedir.

Bu çalışmada da eğitsel veri madenciliğinde sıklıkla kullanılan ve tahmin edici modeller arasında yer alan sınıflandırma teknikleri ve algoritmaları kullanılmış ve sonuçlar irdelenmiştir.

Eğitsel Veri Madenciliği

"Eğitsel Veri Madenciliği", eğitim ortamlarından gelen benzersiz veri türlerini araştırmak için yöntemler geliştirmek ve öğrencileri ve öğrendikleri ortamları daha iyi anlamak için bu yöntemleri kullanmakla ilgilenen "yeni bir disiplin" olarak tanımlanmaktadır [11]. Eğitsel veri madenciliği, bilgisayar bilimi, eğitim ve istatistik alanlarının birleşimi olarak düşünülebilir [12]. Veri madenciliği tekniklerinin eğitim sistemlerine uygulanması, eğitim programlarını sürekli iyileştirmek amacıyla, öğretim tasarımını yeniden biçimlendirmenin bir yolu olarak görülmüştür [13]. Veri madenciliği tekniklerinin eğitim sistemlerini tasarlamak üzere nasıl uygulanacağı Şekil 1'de gösterilmektedir. Şekil 1' de görebileceğimiz gibi, eğitimciler ve akademik sorumlular eğitim sistemlerinin tasarlanması, planlanması, inşa edilmesi ve sürdürülmesinden sorumludur. Keşfedilen bilgi yalnızca eğitim tasarımcıları ve öğretmenler tarafından değil, aynı zamanda kullanıcılar (öğrenciler) tarafından da kullanılabilir. Böylece eğitimciler, öğretim için daha nesnel geri bildirim alabilir, ders içeriğinin yapısını ve öğrenme sürecindeki etkinliğini değerlendirebilirler. Bu değerelendirme sayesinde, öğrenciler rehberlik ve izlemedeki gereksinimlerine göre gruplara ayrılabilir ve öğretimde en sık yapılan hatalar saptanabilir, daha etkili etkinlikler bulunarak, derslere uyarlanması sağlanabilir. Derslerin daha iyi kişiselleştirilmesi için saha yeniden yapılandırabilir, içerik öğrencinin gelişimine göre yeniden düzenlenebilir ve planlanabilir [14]. Bu bağlamda, eğitim sürecini iyileştirmeye yardımcı olacak faydalı bilgileri keşfetmek için farklı veri madenciliği teknikleri uygulanabilir.



Şekil 1. Veri Madenciliği Tekniklerinin Eğitim Sistemlerini Tasarlamak İçin Uygulanması [13].

Eğitsel veri madenciliği yöntemlerinin taksonomisi şu şekildedir [15]:

1.Tahmin
- 1.1. Sınıflandırma
- 1.2. Regresyon
- 1.3. Yoğunluk Hesaplama

2. Kümeleme
3. İlişki Madenciliği
- 3.1. Birliktelik kuralı madenciliği
- 3.2. Korelasyon madenciliği
- 3.3. Sıralı örüntü madenciliği
- 3.4. Nedensel veri madenciliği

4. İnsan yargısı için veri damıtma
5. Modeller ile keşif

Tahmin yönteminde amaç, verilerin başka bir yönünün birleşiminden (öngörücü değişkenler) verilerin tek bir yönünü (öngörülen değişken) çıkaran bir model geliştirmektir. Tahmin yöntemlerinden sınıflandırma (öngörülen değişken kategorik bir değer olduğunda) kullanılır [11]. Bu çalışmada da öğrencilerin akademik başarılarını eğitsel veri madenciliği yöntemleri ile tahmin etmek için sınıflandırma teknikleri ve algoritmaları kullanılmıştır. Bu amaçla, 5., 6., 7. ve 8. sınıf ortaokul öğrencilerinin demografik özellikleri ve sosyoekonomik durumları (öngörücü değişkenler) hakkında elde edilen veriler kullanılarak genel başarı ortalamalarını (öngörülen değişken) tahmin eden bir model geliştirmeye çalışılmıştır. Bu amaçla aşağıdaki sorulara cevap aranmıştır:

1) Öğrencilerin e-okul sistemindeki demografik ve sosyoekonomik verileri kullanılarak oluşturulan farklı sınıflama modelerinin öğrencilerin yıl sonu genel başarı ortalamalarını tahmin etme başarısı nasıldır?

1.1. Öğrencilerin yıl sonu genel başarı ortalamalarının tahmin edilmesinde hangi değişkenler daha önemlidir?

1.2. Performans metrikleri açısından karşılaştırıldığında yıl sonu not ortalaması tahmininde en iyi performans gösteren sınıflama algoritması/algoritmaları nelerdir?

2) Birinci araştırma probleminde en iyi performans gösteren sınıflama algoritması kullanılarak öğrencilerin yıl sonu genel başarı ortalamalarının daha önceden tahmin edilmesi mümkün müdür?

2 Yöntem

Bu araştırmada elde edilen öğrenci verilerinden yıl sonu genel başarı ortalamalarını tahmin etmek amacıyla eğitsel veri madenciliği yöntemlerinden sınıflandırma teknikleri ve algoritmaları kullanılmıştır. Öğrencilerden elde edilen verilere dayanarak, bir öğrencinin notlarını veya diğer öğrenme çıktılarını tahmin etmek için en sık kullanılan teknikler sınıflandırma, kümeleme ve ilişkilendirmedir [16]. Bu çalışmada öğrencilerin yıl sonu genel başarı ortalamalarını tahmininde "Logistic Regression (Lojistik Regresyon), Linear SVM (Doğrusal Destek Vektör Makineleri), Non-Linear SVM (Doğrusal Olmayan Destek Vektör Makineleri), RandomForest (RastgeleOrman), NaiveBayes (Naif Bayes), Bagging, K-nearest neigborhood (K-En Yakın Komşu), Multilayer Perceptron (Yapay Sinir Ağları)" sınıflayıcıları olmak üzere çok çeşitli sınıflandırma yöntemleri denenmiş ve sonuçlar karşılaştırılmıştır. Elde edilen sonuçlara bulgular bölümünde yer verilmiştir.

2.1 Veri Seti

Bu çalışma, 2018-2019 eğitim-öğretim yılı 2. Dönemi'nde, Yalova İlindeki ortaokul öğrencilerinden elde edilen verilerle yapılmıştır. Örneklem seçiminde tabakalı örnekleme yöntemi kullanılmıştır. Tabakalı örnekleme, evrendeki alt grupların belirlenip bunların evren büyüklüğü içindeki oranlarıyla örneklemde temsil edilmelerini amaçlar. Tabakalı örnekleme için önce evren içinde homojen alt gruplar (tabakalar), daha sonra her bir tabaka için alt evren oluşturulur [17]. Bu bağlamda, bu çalışmada sosyo demografik açıdan farklılıklara sahip dört resmi ortaokul örneklem olarak seçilmiştir. Örneklem tabakaları saptandıktan sonra, her tabaka içinde birbirine denk kümeler (her okuldan 5.6.7.8.sınıf öğrencileri) saptanmış ve her küme içinden yine seçkisiz örnekleme yapılarak belirli sayıda öğrenci çekilerek, 1395 örnek seçilmiştir. Öğrencilerin gizli tutularak E-okul Yönetim Bilgi Sistemi'nden elde edilebilen 27 bağımsız değişkenden oluşmaktadır. E-okul Yönetim Bilgi Sistemi'nden elde edilebilen bağımsız değişkenler Tablo 1'de gösterilmiştir.

Tablo 1. E-okul Yönetim Bilgi Sistemi'nden Elde Edilen Bağımsız Değişkenlerin Listesi

Bağımsız Değişkenler	**Açıklama**
Cinsiyet	Öğrenci Cinsiyeti
Yaş	Öğrenci Yaşı
Asağ	Anne Sağ/Ölü Olma Durumu
Bsağ	Baba Sağ/Ölü Olma Durumu
ABayrı	Ebeveynlerin Birlikte/Ayrı Olma Durumu
Ayaşam	Anne İle Yaşama Durumu
Byaşam	Baba İle Yaşama Durumu
AByaşam	Aile İle Yaşama Durumu
Aöğrenim	Anne Eğitim Durumu
Böğrenim	Baban Eğitim Durumu
Açalışma	Anne Çalışma Durumu
Bçalışma	Baba Çalışma Durumu
gelir	Gelir Durumu
Ksayısı	Kardeş Sayısı
oda	Kendine Ait Odası Olma Durumu
hastalık	Sürekli Hastalık Durumu
Öeğitim	Özel Eğitim Durumu
devam	Devamsızlık Durumu
Mkurs	Matematik Kurs Alma Durumu
Tkurs	Türkçe Kurs Alma Durumu
Ykurs	Yabancı Dil Kurs Alma Durumu
Dkurs	Drama Kurs Alma Durumu
Bkurs	Beden Eğitimi Kurs Alma Durumu
Mkurs	Müzik Kurs Alma Durumu
Gkurs	Görsel Sanatlar Kurs Alma Durumu
Ykurs	Yazarlık ve Yazma Becerileri Kurs Alma Durumu
Fkurs	Fen Kurs Alma Durumu

2.2 Öznitelik Seçme

Özellik altkümesi seçimi, olabildiğince alakasız ve gereksiz bilgileri belirleme ve kaldırma işlemidir. Bu işlem, verilerin boyutsallığını azaltır ve öğrenme algoritmalarının daha hızlı ve daha etkili çalışmasına izin verebilir. Özellik alt kümesi seçimi, öğrenme algoritmasının belirli bir problem için önemli özelliklere odaklanmasına yardımcı olabilir [17]. Bu araştırmada da öğrencilerin başarısını etkilediği düşünülen 27 bağımsız değişken bulunmaktadır. Bu bağımsız değişkenlerin, hangilerinin öğrenci başarısını tahmin etmede daha etkili olduğu bilinmemektedir. Bunun için E-okul Yönetim Bilgi Sistemi'nden elde edilebilen bütün bağımsız değişkenler yerine, tahmin sonuçlarını artıracağı düşünülen korelasyona dayalı özellik seçicisi, (Correlation Based Feature Selector - CFS) kullanılarak özellik alt kümesi seçimine gidilmiştir. CFS verilerin daha verimli kullanılması, alakasız verilerin kaldırılması, öğrenme doğruluğunun artırılması amacıyla, makine öğrenimi için bir ön işleme adımı olan korelasyona dayalı özellik seçicisidir [18]. Bu yöntemle özellik alt kümesi seçilirken, sınıfla yüksek oranda korelasyonlu olan, ancak birbiriyle ilişkili olmayan özellikler içeren bir özellik alt kümesi elde edilir [17]. Özellik alt kümesi seçimi ile 27 olan bağımsız değişken sayısı 8'e düşmüştür. CFS uygulanması sonucunda Genel Başarı

Ortalaması (GBO) tahmini için şu bağımsız değişkenler elde edilmiştir: **Özellik Alt Kümesi Seçimi Sonrası Elde Edilen Bağımsız Değişkenler :** yaş, devam, ABayrı, Aöğrenim, Böğrenim, gelir, oda, Fkurs

Araştırmada ayrıca, Temel Bileşen Analizi (Principal Component Analysis- PCA) kullanılmıştır. PCA orijinal veri setindeki bilgilerin çoğunu koruyarak değişken sayısını önemli ölçüde azaltmak için kullanılır. PCA bunu yapmak için boyut küçültme tekniğini kullanarak, verilerin boyutsallığını azaltır [19].

2.3 Sınıflandırma Yöntemleri

Sınıflandırma işlemi en basit şekliyle bağımsız değişken değerleri belli iken, bağımlı değişkenin değerini/düzeyini tahmin etme işlemidir [20]. Gerek istatistik gerekse makine öğrenimi temelli çeşitli sınıflandırma yöntemleri geliştirilmiştir [21].

Bu çalışmada, Logistic Regression (Lojistik Regresyon), Linear SVM (Doğrusal Destek Vektör Makineleri), Non-Linear SVM (Doğrusal Olmayan Destek Vektör Makineleri), RandomForest (RastgeleOrman), NaiveBayes (Naif Bayes), Bagging, K-nearest neigborhood (K-En Yakın Komşu), Multilayer Perceptron (Yapay Sinir Ağları) sınıflandırma yöntemleri kullanılmış olup aşağıda kısaca açıklanmıştır:

NaiveBayes (Naif Bayes), yönteminde Bayes olasılığına bağlı olarak sınıflandırma yapılır. Bayes olasılığı koşulu olasılığın *k* tane ayrık olay için genelleştirilmiş halidir. Bu olasılık aşağıdaki şekilde tanımlanır:

$$P(C_J / X) = P(X / C_J)\, P(C_J) / P(X) \qquad (1)$$

P (C / X_J): X durumu verilmişken C_J sınıfının ortaya çıkma olasılığı

P (X / C_J): C_J sınıfında *X* durumunun ortaya çıkma olasılığı
P (C_J): (C_J) sınıfının ortaya çıkma olasılığı
P (X): X durumunun ortaya çıkma olasılığı.

Bayes sınıflamada amaç, *X = (X_1.....Xp)* yani bağımsız değişken vektörünün değeri biliniyorken bağımlı değişken değerini tahmin etmektir. Bağımlı değişken değerini tahmin etmek için *P (C_J / X)* şeklindeki Bayes olasılıkları hesaplanarak en büyük olasılık değerine ait sınıf seçilir [22].

K-nearest neigborhood (K-En Yakın Komşu), eğitim setinde test nesnesine en yakın olan bir grup k nesnesi bulur ve bu k nesnesine belli bir sınıfın atanmasını temel alır. Bu yaklaşımın üç temel unsuru vardır: Bilinen nesneler kümesi, nesneler arasındaki mesafeyi hesaplamak için benzerlik ölçümü ve en yakın komşuların sayısı olan k'nın değeri. Bilinmeyen bir nesneyi sınıflandırmak için, bu nesnenin bilinen nesnelere olan mesafesi hesaplanır, en yakın k komşuları tanımlanır ve bu en yakın komşuların sınıf etiketleri daha sonra nesnenin sınıf etiketini belirlemek için kullanılır. Bir eğitim seti D ve bir test nesnesi x = (x′, y′) verildiğinde, algoritma z ile tüm eğitim nesneleri (x, y) ∈ D arasındaki mesafeyi (veya benzerliğini) en yakın komşu listesini belirlemek için hesaplar, D***z*** (x, bir eğitim nesnesinin verileri, y ise onun sınıfıdır. Benzer şekilde, x′ test nesnesinin verileri ve y′ ise onun sınıfıdır). En yakın komşu listesi elde edildikten, test nesnesi en yakın komşusunun çoğunluk sınıfına göre sınıflandırılır [23].

SVM (Destek Vektör Makineleri), tüm bilinen algoritmalar arasında en sağlam ve doğru yöntemlerden birini sunar. SVM'nin amacı, eğitim verilerindeki iki sınıfın üyelerini ayırt etmek için en iyi sınıflandırma işlevini bulmaktır. Doğrusal olarak ayrılabilir bir veri kümesi için, doğrusal bir sınıflandırma işlevi sağlar ve iki sınıfı biribirinden ayıran bir hiper düzleme karşılık gelir. İki sınıf arasındaki aralığı maksimize ederek, en uygun yüksek boyutlu bir hiper düzlemi bulur [24].

RandomForest (RastgeleOrman) yöntemi, sınıflandırma amacıyla kullanılan bir başka topluluk öğrenme yöntemidir. RastgeleOrman, her biri birbirinden bağımsız olarak ve aynı dağılım kullanılarak eğitim verisinden rastgele elde edilmiş bir örnekleme dayanan karar ağaçlarından oluşturulan bir topluluktur. Bu yöntem eğitim sırasında birçok karar ağacı oluşturur ve daha sonra kestirim sırasında bu karar ağaçlarının sınıflandırma sonuçlarından yararlanılarak, girdinin sınıfına göre çoğunluk oyu aracılığıyla karar belirlenir [25].

Bagging yöntemi, orijinal veri setinden elde edilen bootstrap örneklerine tahminciler uygulanarak bir topluluk oluşturur. Bu arada bootstrap uygulaması, iadeli rasgele seçim yapıp alt örneklemler oluşturmak için kullanılır. Orijinal veri setindeki sayı ile aynı olacak alt örneklemler oluşturur. Bu nedenle bazı gözlemler bootstrap sonucunda oluşturulan örneklemlerde yer almazken bazıları iki veya daha fazla defa görülebilir. Tahminlerin birleştirilmesi aşamasında sınıflandırma ağaçlarında sonuçlar oylama ile belirlenir [26].

$$\eta b\,(X \mid Z1), \ldots, \eta\,(X \mid Zb) \qquad (2)$$ formülü kullanılır [26].

Multilayer Perceptron (Yapay Sinir Ağları) nöronlardan oluşur. Ağ içerisindeki bir nöron diğer nöronlara sinyaller gönderir, böylece gelen girdiler tanımlanır. Bir Y nöronu ele alınırsa, bu nöron X_1, X_2, X_3 nöronlarından işaret alır. Daha sonra X_1, X_2, X_3 nöronlarını Y nöronuna bağlayan ağırlıklar (w_1, w_2, w_3) hesaplanır. Öğrenme sürecinde verilerin çıktı katmanına ulaşabilmesi için w ağırlıkların hesaplanması gerekir. Öğrenme için ayrılmış veri kümesi üzerinde bu ağırlıklar hesaplandıktan sonra, diğer veri kümesi ile de öğrenmenin ne kadar gerçekleştiğini bulmak için ağırlıklar test edilir. Test işlemi sonunda ağırlıkların etkinliği doğrulanırsa, öğrenme işlemi tamamlanır [27]. Formülle tanımlanacak olursa bir *Y-girdi* nöronu gelen sinyallerin ağırlıklarla çarpımının toplamıdır:

$$Y\text{-}girdi = w_1x_1 + w_2x_2 + w_3x_3 \qquad (3)$$

Logistic (Lojistik), popüler bir regresyon yöntemidir. Lojistik modelin dayandığı matematiksel form f (z) olarak tanımlanır. f (z) işlevi 0 ile 1 arasında değişir. Model, her zaman 0 ile 1 arasında bir sayı olan bir olasılığı tanımlamak için tasarlanmıştır.

$$\text{Range: } 0 \le f(z) \le 1 \qquad (4)$$

z değeri - ∞ olduğunda; f (z) lojistik fonksiyonu 0'a eşit olur. Z değeri +∞ olduğunda; f (z) lojistik fonksiyonu 0'a eşit olur. Lojistik model, asla 1'in üstünde veya 0'ın altında bir risk tahmini almaz. Bu, diğer olası modeller için her zaman doğru değildir, bu yüzden bir olasılık tahmin edildiğinde lojistik model genellikle ilk tercihtir [28**].**

3 Bulgular

Bu çalışmada, 5, 6, 7 ve 8. ortaokul öğrencilerinin dönem sonu genel başarı ortalamaları, öğrencilerin sosyoekonomik ve demografik özelliklerine dair verileri kullanılarak sınıflandırma yöntemleri ile tahmin edilmiştir. Öğrencilerin yılsonu notları

Tablo 2'de görülen Milli Eğitim Bakanlığı ilköğretim not ölçeğine göre sınıflandırma teknikleri ve algoritmaları ile tahmin edilmiştir. Bu çalışmada dokuz sınıflandırıcı algoritma ile sınıflandırma testleri gerçekleştirilmiştir. Analizler yapılırken, sınıflandırıcı algoritmaların yalnız kullanılmasının (standalone) yanısıra çoklu meta sınıflandırıcı (MultiClassClasifier) algoritmaları kullanılarak da analizler yapılmıştır. (MultiClassClasifier) Çoklu Sınıflandırıcı Algoritmalara ait sonuçlara bulgularda yer verilmiştir.

Tablo 2. MEB İlköğretim Not Ölçeği

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

Bu araştırmada, sınıflandırma teknik ve algoritmaları uygulanırken, algoritmanın performansını değerlendirmek, en iyi sonuca ulaşabilmek için, 10 - kat çapraz doğrulama (Cross-validation) test tekniği kullanılmıştır. Cross-validation (CV) test tekniğinde, verilerin bir kısmı (eğitim örneği) algoritmayı eğitmek için kullanılırken ve geri kalan veriler (doğrulama örneği) olarak kullanılır. Çapraz doğrulama, tek bir veri setinden eğitim ve test setlerinin bir dağılımını oluşturur. Çapraz doğrulama işleminde veriler, her biri kat olarak adlandırılan k altkümelerine S1..... Sk ayrılır. Daha sonra öğrenme algoritması, eğitim seti olarak Si dışındaki tüm altkümelerin birleşimini ve test seti olarak Si'yi kullanarak, her seferinde, i = 1 ila k için k kere uygulanır [29]. Waikato Bilgi Analizi Ortamı (WEKA) programında var olan tüm algoritmalar bu veri dosyası üzerinde sırayla çalıştırılmış ve en yüksek korelasyon katsayısı veren dokuz algoritma seçilerek tablolaştırılmıştır.

Tablo 3. En İyi Performans Gösteren Sınıflandırıcı Algoritmalar

Sınıflandırıcı Algoritma	Doğruluk (%)
Logistic	**64.00**
Naïve Bayes	60.60
Linear SVM	62.20
(k-NN) (k=10)	**59.30**
RandomForest	61.60
Non-Linear SVM	63.05
Bagging	61.90
Multilayer Perceptron	60.40

Tablo 3'te görüldüğü üzere, sınıflandırıcı algoritmaların yalnız uygulanmaları sonucunda, genel başarı ortalaması tahmininde logistic (%64.00), Linear SVM (%62.20), Non-Linear SVM (%63.05) algoritmaları en iyi sonucu vermiştir. En düşük başarı tahmini k-Nearest Neighborhood (k-NN) algoritmasında (% 59.30) görülürken en iyi başarı tahmini *logistic* algoritmasında olmuştur. Genel olarak sınıflandırıcılar arasında çok büyük başarı farklılıkları yoktur. Önişlem sürecinde korelasyona dayalı özellik seçici (CFS) uygulanması sonucunda 27 olan bağımsız değişken sayısı 8'e düşmüştür. Genel başarı ortalaması tahminine elde edilen bağımsız değişkenlerle devam edilmiştir (yaş, devam, ABayrı, Aöğrenim, Böğrenim, gelir, oda, Fkurs).

Tablo 4. Korelasyona Dayalı Özellik Seçici Uygulanması İle Seçilen Bağımsız Değişkenlere Göre Sınıflandırıcı Algoritmaların Tahmin Sonuçları

Sınıflandırıcı Algoritma	Doğruluk (%)
Logistic	**63.50**
Naïve Bayes	60.90
Linear SVM	63.00
k-NN (k=10)	60.01
RandomForest	**58.10**
Non-Linear SVM	61.40
Bagging	60.20
Multilayer	60.80

Tablo 4'te görüldüğü üzere, CFS uygulanması sonucu k-NN algoritması ile genel başarı ortalaması tahmininde başarı oldukça artmıştır. Naive Bayes ve Multilayer Perceptron yöntemlerinde öznitelik seçme ile başarımda bir miktar artış gözlenmiş olup diğer yöntemlerde başarımın düşmesine yol açmıştır. CFS uygulanması sonucunda GBO'sında en başarılı tahmin yine *logistic* algoritmasında olmuştur. En düşük tahmin değeri RandomForest (% 58.10) algoritmasında olmuştur. MultiClassClassifier, kullanılarak en iyi başarı tahmini veren 8 algoritma tekrar denenerek analizlere devam edilmiştir.

Tablo 5. MultiClassClassifier ile Yapılan Sınıflandırıcı Algoritmaların Tahmin Sonuçları

Sınıflandırıcı Algoritma	**Doğruluk (%)**
Logistic	**62.15**
Naïve Bayes	60.45
Linear SVM	56.20
k-NN (k=30)	60.16
RandomForest	61.03
Non-Linear SVM	**55.81**
Bagging	59.94
Multilayer Perceptron	58.54

Tablo 5'te görüldüğü üzere, MultiClassClassifier kullanılarak yapılan sınıflamada algoritmaların tahmin sonuçlarında çok büyük başarı farklılıkları görülmemiştir. Sınıflandırıcı algoritmaların yalnız uygulanması ile kıyaslandığında, MultiClassClassifier ile yapılan sınıflamada sadece k-NN (%60.10) algoritmasında artış olduğu görülmüş, diğer algoritmalarda GBO tahmininde düşüş görülmüştür. En düşük başarı tahmini veren algoritma, Non-Linear SVM olmuştur. En yüksek tahmin veren algoritma yine logistic olmuştur. CFS ile MultiClassClassifier ile yapılan sınıflama analizleri karşılaştırıldığında, çoklu sınıflandırmada yine sadece *k-NN* (%60.10) algoritmasında bir miktar artış olmuş, diğer algoritmalarda başarı tahmininde düşüş yaşanmıştır. Bu bağlamda, CFS ile yapılan analiz sonuçlarının MultiClassClassifier ile yapılan sınıflama analizlerinden daha iyi tahmin sonuçları verdiği söylenebilir. MultiClassClassifier ile CFS birlikte kullanılarak analizler 8 algoritma için tekrar edilmiştir. Tablo 6'da, MultiClassClassifier ile CFS ile yapılan sınıflandırıcı algoritmaların genel başarı ortalaması tahmin sonuçları görülmektedir.

Tablo 6'da görüldüğü üzere, MultiClassClassifier ile CFS birlikte kullanılarak yapılan sınıflandırıcı algoritmaların tahmin sonuçları ile sınıflandırıcı algoritmaların yalnız kullanımından elde edilen tahmin sonuçları karşılaştırıldığında, sadece Naive Bayes (% 61.04), k-NN (% 61.34) algoritmalarında genel başarı ortalaması tahmininde bir miktar artış sağlanmış, diğer algoritmalarda başarı tahmininde Multilayer algoritması dışında düşüş görülmüştür. MultiClassClassifier ile CFS birlikte kullanılarak yapılan sınıflandırıcı algoritmaların tahmin sonuçları ile CFS yalnız kullanımından elde edilen sonuçlar karşılaştırıldığında, başarım sadece, Naïve Bayes (%61.04), k-NN (% 61.34), RandomForest (% 58.61) algoritmalarında görülmüştür. MultiClassClassifier ile CFS birlikte kullanılarak yapılan sınıflandırıcı algoritmaların tahmin sonuçlarında yine en yüksek genel başarı ortalaması tahmini *logistic* algoritmasında, en düşük başarı tahmini, Non-Linear SVM algoritmasında olmuştur. Bu bağlamda, CFS kullanılarak yapılan sınıflandırıcı algoritmaların, MultiClassClassifier ile CFS birlikte kullanılarak yapılan analiz sonuçlarına göre daha başarılı tahmin sonuçları verdiği söylenebilir. PCA kullanılarak 27 olan değişken sayısı 19 adet değişken örüntüsüne dönüşmüştür. Elde edilen 19 adet değişken örüntüsü üzerinde sınıflandırıcı algoritmalar kullanılarak analizlere devam edilmiştir.

Tablo 6. MultiClassClassifier ve CFS ile Yapılan Sınıflandırıcı Algoritmaların Tahmin Sonuçları

Sınıflandırıcı Algoritma	**Doğruluk (%)**
Logistic	**62.10**
Naïve Bayes	61.04
Linear SVM	56.11
k-NN (k=18)	61.34
RandomForest	58.61
Non-Linear SVM	**56.03**
Bagging	60.90
Multilayer Perceptron	60.40

Tablo 7'de görüldüğü üzere, PCA ile seçilen özniteliklere göre sınıflandırıcı algoritmaların doğruluk sonuçları, CFS başarımı ile kıyaslandığında, logistic, k-NN, RandomForest, Linear SVM, Multilayer Perceptron algoritmalarında genel başarı ortalama tahmininde artış görülmüştür. Fakat yine Genel Başarı Ortalaması tahmininde sınıflandırıcı algoritmalar arasında çok büyük başarı farklılıkları yoktur. Sınıflandırıcı algoritmaların yalnız uygulanması başarımı ile PCA ile seçilen özniteliklere göre sınıflandırıcı algoritmaların tahmin sonuçları karşılaştırıldığında, yine algoritmalar arasında çok büyük başarı farklılıkları gözlenmemiştir. Logistic algoritması, şimdiye kadar yapılan analizlerin tümünde en yüksek tahmin değerini veren algoritma olmasına rağmen, PCA kullanılarak yapılan analiz sonucunda % 64.13 tahmin değeri ile *en yüksek değere* ulaşmıştır.

Tablo 7. PCA Kullanılarak Oluşturulan Özniteliklere Göre Sınıflandırıcı Algoritmaların Tahmin Sonuçları

Sınıflandırıcı Algoritma	**Doğruluk (%)**
Logistic	**64.13**
Naïve Bayes	**54.86**
Linear SVM	62.66
k-NN (k=30)	61.20
RandomForest	59.50

Sınıflandırıcı Algoritma	Doğruluk (%)
Non-Linear SVM	62.96
Bagging	59.20
Multilayer Perceptron	62.33

Tablo 8'de görüldüğü üzere, Genel başarı ortalaması tahmininde PCA yapıldığında, Logistic ve Multilayer Perceptron algoritmaları; hem CFS hem de MultiClassClassifier ile birlikte uygulandığında, Naive Bayes ve k-NN algoritmaları; CFS tek başına uygulandığında; Linear SVM; sınıflandırıcı algoritmaların yalnız uygulanmaları sonucunda; RandomForest, Non-Linear SVM, Bagging algoritmaları en iyi tahmini gerçekleştirmiştir. GBO'nda sınıflandırıcı algoritmaların yalnız uygulanması; CFS ile uygulanması, PCA'nın yalnız uygulanması, MultiClassClassifier yalnız uygulanması, CFS - MultiClassClassifier yöntemlerinin birlikte uygulanması sonucu genel başarı ortalaması başarımında *logistic* algoritması en iyi tahmini gerçekleştirmiştir

Tablo 8. Eğitim Verileri Üzerinde Yapılan Tüm Sınıflandırıcı Algoritmaların Tahmin Sonuçları

Sınıflandırıcı Algoritma	Temel Bileşenler Analizi-Doğruluk (%)	Korealasyona Dayalı Özellik Seçicisi Doğruluk (%)	Tekli Meta Sınıflandırıcı Doğruluk (%)	Çoklu Meta Sınıflandırıcı Doğruluk (%)	Çoklu Meta Sınıflandırıcı Korealasyona Dayalı Özellik Seçicisi Doğruluk (%)
Logistic	**64.13**	63.50	64.00	62.15	62.10
Naïve Bayes	54.86	60.90	60.60	60.45	**61.04**
Linear SVM	62.66	**63.00**	62.20	56.20	56.11
k-NN (k=30)	61.20	60.01	59.30	60.16	**61.34**
RandomForest	59.50	58.10	**61.60**	61.03	58.61
Non-Linear SVM	62.96	61.40	**63.05**	55.81	56.03
Bagging	59.20	60.20	**61.90**	59.94	60.90
Multilayer Perceptron	**62.33**	60.80	60.40	58.54	60.40

4 Sonuçlar ve Öneriler

Eğitsel Veri Madenciliği, eğitim ortamlarından gelen benzersiz veri türlerini araştırmak için yöntemler geliştiren, öğrencileri ve öğrendikleri ortamları daha iyi anlamak için bu yöntemleri kullananan yeni bir disiplindir. Literatüre göre, elde edilen veriler kullanılarak bu verilerden faydalı bilgilere ulaşmak için kullanılan veri madenciliğinin eğitimde kullanılmasının da eğitimcilere yol göstermesi açısından faydalı bilgiler sağlayacağı söylenebilir. Bu amaçla bu araştırmada da, ortaokul 5, 6, 7 ve 8. sınıf öğrencilerinin, demografik ve sosyoekonomik özelliklerinin akademik başarılarına olan etkilerini anlamak için E-okul Yönetim Bilgi sisteminden elde edilen verileri (27 bağımsız değişken) kullanılmıştır. Sonrasında elde edilen verilerden öğrencilerin dönem sonu genel başarı ortalamalarını tahmin etmek için, sınıflandırma yöntemleri ve algoritmaları kullanılmıştır. WEKA programında var olan tüm algoritmalar bu veri dosyası üzerinde sırayla çalıştırılmış ve en yüksek korelasyon katsayısı veren sekiz algoritma seçilerek değerlendirilme yapılmıştır. Deneysel sonuçlara göre, genel başarı ortalaması tahmininde sınıflandırma yöntemlerinde başarılı sonuçlar elde edilmiştir. Genel başarı ortalaması tahmininde *PCA* kullanılarak yapılan analiz de *logistic* sınıflandırma algoritması en iyi başarımı göstermiştir. Çoklu meta sınıflandırıcı, algoritmaların yalnız uygulanması durumunda genel başarı ortalamada sadece *k-NN* de artış sağlanmıştır. CFS kullanılarak, 27 olan öznitelik sayısı, 8'e düşürülerek analizler tekrarlandığında, Naive Bayes, k-NN, Linear SVM, Multilayer Perceptron sınıflandırma algoritmalarında bir miktar artış gözlenmiştir. Çoklu sınıflandırıcı ile yapılan analizlerle CFS kullanılarak yapılan analiz sonuçları karşılaştırıldığında, öznitelik seçme yönteminde GBO tahmininde daha fazla başarı sağlanmıştır. Öznitelik seçme yöntemi ve çoklu sınıflandırıcı kullanılarak analizler tekrarlandığında, sınıflandırıcıların yalnız kullanımına göre tahmin sonuçlarında artış yaşanmamıştır. Yine GBO'nda sınıflandırıcı algoritmaların yalnız uygulanması, öznitelik seçme yöntemi ile uygulanması, Temel Bileşenler Analizi ile uygulanması, Çoklu Meta Sınıflandırıcı, Korelasyona Dayalı Özellik Seçicisinin birlikte uygulanması sonucu GB ortalaması başarımında *logistic* algoritması en iyi tahmini gerçekleştirmiştir. Bu çalışmada akademik başarıyı etkilediği düşünülen ve E-okul Yönetim Bilgi Sistemi'nden elde edilebilen veriler ile araştırma yapılmıştır. Gelecekteki araştırmalar için akademik başarıyı etkilediği düşünülen diğer faktörler (öğrencinin sürekli katıldığı sosyal etkinlikler, evde toplam ders çalışma süresi, oyun oynama sıklığı vb.) faktörlerde araştırmaya dahil edilerek öğrencilere ve ebeveynlerine çevrimiçi veya çevrimdışı uygulanabilecek anketler ile dönem sonu herhangi bir ders ortalaması veya genel başarı ortalamasına yönelik başarım tahmini çalışması yapılabilir.

Veri Erişebilirliği

Bu çalışmada kullanılan veri seti ve .arff dosyasına aşağıdaki bağlantılardan ulaşabilirsiniz.

https://drive.google.com/file/d/11eIZm_oIuFmMIJaG-wNCCzGpgOhDWPLYP/view?usp=sharing

https://drive.google.com/file/d/1HVwD94dBx7JPTb4LFhgwlnlGzIt-X0NR/view?usp=sharing

Kaynaklar

1. Bourdieu P. "Culture reproduction and social reproduction," in Knowledge, Education, and Cultural Change, Editor: Brown R. London, Tavistock, 1973.
2. Bourdieu P, Passeron JC. "Reproduction in Education, Society and Culture", Vol. 4, Newbury Park, CA: Sage, 1990.
3. Coleman, J. S. "Social capital in the creation of human capital". Am.J. Sociol. 94,S95–S120. doi:10.1086/228943, 1988.
4. Pettigrew EJ. "A Study of the impact of scioeconomic status on student achievement in a rural east Tennessee school system". Electronic Theses and Dissertations. Paper 1844, 2009.
5. Akhtar Z, Niazi K. "The relationship between socio-economic status and learning achievement of students at secondary level". International Journal of Academic Research, 3(2), 956-961, 2011.
6. Gelbal S. "The effect of socio-economic status of eighth grade students on their achievement". Turkish Education and Science, 33(150), 1-13, 2008.
7. Şirin SR. "Socioeconomic status and academic achievement: A meta-analytic review of research". Review of Educational Research,75, 417–453,2005.
8. Peña-Ayala A."Educational Data Mining: A survey and a data mining-based analysis of recent works". Expert systems with applications, 41(4), 1432-1462, 2014.
9. Gök M. "Makine Öğrenmesi Yöntemleri ile Akademik Başarının Tahmin Edilmesi". Gazi Üniversitesi Fen Bilimleri Dergisi, Part C, Tasarım Ve Teknoloji, GU J Sci, Part C, 5(3):139-148, 2017.
10. Petcu N. "Data mining techniques used to analyze students opinions about computization in the educational system". Bulletin of the Transilvania University of Brasov. Economic Sciences. Series V, 8(1), 289, 2015.
11. Bousbia N, Belamri I. Which Contribution Does EDM Provide to Computer-Based Learning Environments? Editor: Peña-Ayala A, Educational data mining (s.3-25). Volume, 524, Newyork, Springer, 2014.
12. Peña-Ayala A."Educational Data Mining: A survey and a data mining-based analysis of recent works". Expert systems with applications, 41(4), 1432-1462, 2014.
13. Romero C, Ventura S, Pechenizkiy M, Baker Ryan SJ d. Handbook of Educational Data Mining. Chapman, Hall/CRC Data Mining and Knowledge Discovery Series, CRC Press, 2011.
14. Romero C, Ventura S. "Educational data mining: a survey from 1995 to 2005". Expert System with Applications, 33, 135-146, 2007.
15. Baker RSJD, Yacef K. "The State of Educational Data Mining in 2009: A Review and Future Visions. Journal of Educational Data Mining", Article 1, Vol 1, No 1, Fall 2009.
16. Hämäläinen W, Vinni M. Classifiers for Educational Data Mining. Editors: Romero C, Ventura, S Pechenizkiy, M Baker, RSJD. Handbook of Educational Data Mining, 2011.
17. Akgün ÖE, Büyüköztürk Ş, Çakmak EK, Demirel F, Karadeniz Ş. Bilimsel Araştırma Yöntemleri. 3. Bölüm. Örnekleme Yöntemleri. 22. Baskı. Ankara: Pegem Akademi, 2008.
18. Hall MA. "Correlation-based Feature Selection for Machine Learning". Doctoral dissertation, University of

Waikato, Dept. of Computer Science, Hamilton, NewZaland, 1999.
19. Yu L, Liu H. "Feature Selection for High-Dimensional Data: A Fast Correlation-Based Filter Solution". Department of Computer Science & Engineering, Arizona State University, Tempe, AZ 85287-5406, USA, 2003.
20. Jolliffe I. Principal Component Analysis.Editor: Lovric M. International Encyclopedia of Statistical Science. Springer, Berlin, Heidelberg, 2014.
21. Zaki MJ, Wagner M. Data Mining and Analysis: Fundamental Concepts and Algorithms. Online (and Offline) Robust PCA, Novel, 2013.
22. Gürsoy T. Veri Madenciliğinde Güncel Yaklaşımlar. 1. Baskı, Çağlayan Yayıncılık, İstanbul, 2014.
23. Altunkaynak B. Veri Madenciliği Yöntemleri ve R Uygulamaları Kavramlar-Modeller-Algoritmalar. 1. Baskı. Seçkin Yayıncılık, Ankara, 2017.
24. Wu X, Kumar V, Quinlan JR, Ghosh J, Yang O, Motoda H, McLachlan GJ, Liu B, Yu PS, Zhou Z, Steinbach, M, Hand DJ, Steinberg D, 2007. "10 Algorithms in Data Mining". *Knowledge & Information Systems*, Jan 2008, Vol. 14 Issue 1, p1-37, 37p, 4 Diagrams, 2 Graphs; DOI: 10.1007/s10115-007-0114-2.
25. Breiman L. "RandomForests". Machine Learning, 45(1), 5-32, 2001.
26. Gelbal S. "The effect of socio-economic status of eighth grade students on their achievement". Turkish Education and Science, 33(150), 1-13, 2008.
27. Steele MB. "Exact bootstrap k-nearest neighbor learners". Mach Learn, 2009.74:235–255 DOI 10.1007/s10994-008-5096-0.
28. Kleinbaum DG, Klein M. Logistic Regression. A Self-Learning Text.Third Edition, Springer New York, 2010.
29. Sammut C, Webb G. "Encyclopedia of Machine Learning. Cross Validation", 2010.