



Veri Madenciliğinde Kullanılan Öğrenme Yöntemlerinin Farklı Koşullar Altında Karşılaştırılması¹

MAKALE TÜRÜ	Başvuru Tarihi	Kabul Tarihi	Erken Görünüm Tarihi
Araştırma Makalesi	26.9.2018	10.11.2018	12.11.2018

Gökhan Aksu ²
Adnan Menderes Üniversitesi

Nuri Doğan ³
Hacettepe Üniversitesi

Öz

Bu çalışmada veri madenciliği ve makine öğrenme yaklaşımının eğitim alanında kullanılması ve bu algoritmalara dayalı olarak elde edilen sonuçların güvenilirlik ve geçerlik değerlerinin ne düzeyde olduğu belirlenmeye çalışılmıştır. PISA 2015 Türkiye ortalamasına göre öğrencilerin başarılı ve başarısız olarak sınıflandığı çalışmada farklı öğrenme yöntemleri kullanılarak fen okuryazarlığı bakımından öğrencilerin hangi sınıfta yer alacağı tahmin edilmiş ve bu aşamada elde edilen sonuçların güvenilirlik ve geçerlik ölçütleri incelenmiştir. Çalışma kapsamında ele alınan 8 farklı öğrenme yönteminden doğru sınıflama sayısı, doğru sınıflama oranı, kappa istatistiği, karekök hata ve göreceli karekök hata değerleri bakımından en iyi sonuçların Random Forest yöntemiyle elde edilirken Ridge lojistik regresyon, Lojistik model ve Hoefding tree yöntemlerinin en başarılı diğer yöntemler olduğu belirlenmiştir. Çapraz geçirme yöntemi kullanılmadan tüm veri setinin eğitim ve test veri seti olarak ayrılması durumunda Lojistik model, Random Forest ve Ridge Regresyon yöntemlerinin farklı büyüklükteki test verilerinde en düşük hata değerlerini verirken Random Tree ve J.48 yönteminin en yüksek hata değerlerine sahip olduğu belirlenmiştir. Ridge regresyon, Random forest ve Lojistik model ile elde edilen hata değerlerinin de farklı yüzdelikteki test verilerinde oldukça tutarlı olduğu sonucuna ulaşılmıştır. Farklı yöntemler yardımıyla elde edilen ölçme sonuçlarının veri setini test ve eğitim verisi olarak ayırmayıp aynı veri seti üzerinden hem öğrenme yöntemini eğitip hem de test ettiğimiz takdirde özellikle Random tree ve J.48 öğrenme yöntemlerinin gerçek performanslarından daha yüksek doğru sınıflama oranına sahip oldukları belirlenmiştir.

Anahtar sözcükler: Veri madenciliği, WEKA, öğrenme yöntemi, sınıflama, PISA.

¹Bu çalışma Aksu (2018) tarafından yapılan “PISA Başarısını Tahmin Etmede Kullanılan Veri Madenciliği Yöntemlerinin İncelenmesi” isimli doktora tez çalışmasından türetilmiştir.

²*Sorumlu Yazar:* Dr., Adnan Menderes Üniversitesi, Aydın Meslek Yüksekokulu, E-posta: gokhanaksu1983@hotmail.com, <https://orcid.org/0000-0003-2563-6112>

³Prof. Dr., Hacettepe Üniversitesi Eğitim Fakültesi, Eğitim Bilimleri Bölümü, E-posta: nuridogan2004@gmail.com, <https://orcid.org/0000-0001-6274-2016>

Günümüzde bireyler hakkında farklı ortamlarda ve farklı amaçlarla çok fazla bilgi toplanmaktadır. Bu bilgiler yardımıyla bireylerin birçok farklı özelliğine dayalı olarak onların satın alma, yeme-içme ve benzeri davranışları hakkında tahminde bulunulmaktadır (North, 2012). Bu aşamada elde edilen çok fazla veri arasından hangilerinin karar verme sürecinde anlamlı, hangilerinin değersiz olduğunun belirlenmesi büyük önem taşımaktadır. Bireylerin özelliklerine ilişkin değişkenlerin yapılan kestirim ve karar alma işlemlerinde elde edilen sonuçların doğruluğu üzerinde büyük etkisi bulunduğundan hangi değişkenlerin bizim için önemli olduğunun belirlenmesi de gerekmektedir (Brown, 2014).

Bilgisayar yazılımlarının gelişimiyle çok fazla veri arasındaki gizli örüntüleri ortaya çıkarmak ve geleceğe ilişkin tahminde bulunmak araştırmacıların ve karar alıcı mercilerin üzerinde çalıştıkları bir konudur. Büyük verilerden anlamlı sonuçlar çıkarabilmek için veri madenciliği (VM) yöntemleri ile bireylerin davranış örüntüleri analiz edilerek gelecekteki davranışları hakkında tahminde bulunulmaktadır (Mavroforakis, 2011). Sınıflandırma ve tahminde bulunma amacıyla kullanılan veri madenciliği yöntemleri, birçok bilim dalında yaygınlaşırken eğitim alanında çok fazla kullanılmadığı görülmektedir (Sinharay, 2016). Denetimli öğrenme yöntemi olarak tanımlanan VM yöntemlerinin klasik yöntemlere göre daha iyi bir kestirim gücüne ve daha az soruna sahip olduğu belirtilmektedir (Fernandez-Delgado, Cernadas, Barro ve Amorim, 2014).

Veri tabanlarından elde edilen bilginin keşfi sürecinde verilerin ve yapılan işin özelliklerinin bilinmemesi durumunda, VM algoritması ne kadar iyi olursa olsun yararlı sonuçlar vermesi olanaklı değildir. Dolayısıyla ilk önce verilerin ve işin özelliklerinin öğrenilmesi gerekir. Bu nedenle sırasıyla: Problemin Tanımlanması, Verilerin Hazırlanması, Modelin Kurulması ve Değerlendirilmesi, Modelin Kullanılması ve Modelin İzlenmesi veri tabanlarında bilgi keşfi sürecinde izlenmesi gereken temel aşamalardır (Chen ve Liu, 1997).

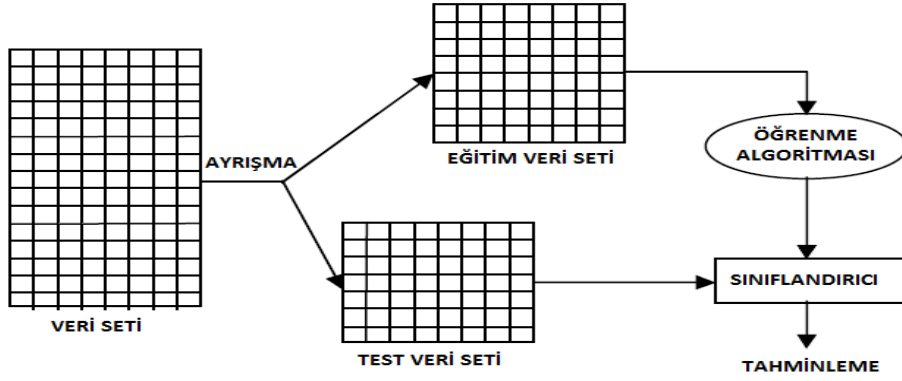
Bilgi çağının yaşandığı günümüzde çok çeşitli kaynaklardan veri toplanmaktadır. Ticari veya bilimsel amaçlı olarak giderek artan verinin görece olarak düşük maliyetle depolanmasını olanak kılan teknolojilerin yanı sıra, bu tür verilerin içinde bulunabilen bir bilginin doğru bir şekilde kullanılması büyük önem taşımaktadır (Bramer, 2013). Belirli bir veri seti üzerinden elde edilen sınıflandırıcının performansını tahmin etmek için kullanılacak en iyi ölçüt tahminin doğruluğu ya da doğru sınıflandırma oranıdır. Her ne kadar doğruluk oranı en önemli ölçüt olarak görülse de sınıflandırıcının elde ettiği sonuçları değerlendirmek için başka ölçütler de bulunmaktadır (Bramer, 2013).

Model değerlendirmesinde kullanılacak yöntemler temel olarak 4 başlık altında yer almaktadır. Bunlar: Bekletme (hold-out), K-katlı çapraz geçirme (k-fold cross validation), Bir tanesini kapsam dışı bırakma (Leave-one-out) ve Bootstrap yöntemidir. Bekletme yaklaşımı aslında tüm verileri eğitim ve test verisi şeklinde belli oranlarda ayırıştırma işlemidir (Souza, Matwin ve Japkowicz, 2002). Bu aşamada veri kümesinden test için belirli bir miktar (% 20 - % 30) tutulur ve bu işleme, bekletme

prosedürü denilir. Daha sonra geriye kalan veriler eğitim için kullanılır. Pratik olarak eldeki verinin üçte birini (1/3) test için, geri kalan kısmını ise eğitim için kullanmak oldukça yaygın bir yöntemdir. Ancak eğer gerekli ise eğitim için kullanılacak verinin bir kısmı doğrulama verisi olarak ayrılabilir (Ahmed ve Elaraby, 2014).

Verileri Eğitime ve Test Etme

Bu yöntemde elinizdeki veri dosyası eğitim ve test verileri olarak ikiye ayrılmaktadır. İlk olarak sınıflandırıcı (classifier) olarak isimlendirilen öğrenme yöntemini oluşturmak için eğitim veri seti oluşturulur. Daha sonra test veri setinden oluşturmuş olduğunuz sınıflandırıcının örneklerin sınıflarını tahmin etmesi sağlanır. Eğer toplam N örnekten oluşan veri setinizin C tanesi doğru olarak sınıflanmış ise bu durumda tahmin edilen doğruluk oranı $p=C/N$ olacaktır. Bu formül hangi sınıfta olduğunuzu bilmediğiniz veriler için kullanılacak sınıflandırıcının performans ölçütlerinden biri olan tahmin edilen doğruluk oranı olacaktır. Görsel olarak eğitim ve test sürecinin nasıl çalıştığı Şekil 1’de gösterilmiştir (Bramer, 2013).

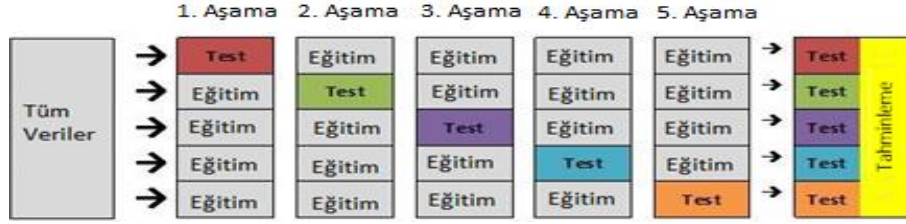


Şekil 1. Eğitim ve test verileri üzerinden öğrenme süreci

Çapraz Geçerlik

Veri madenciliği için kullanılacak olan verilerin az ve sınırlı olması durumunda elde edilen sonuçların değerlendirilmesi için bekletme yaklaşımına seçenек olarak kullanabileceğiniz değerlendirme yöntemi k-katlı çapraz geçerleme yaklaşımıdır. Toplam N örnekten oluşan veri seti k adet (5 veya 10 gibi) eşit parçaya ayrılır (Eğer toplam örnek sayısı k sayısına tam olarak bölünemiyorsa ayrılan son parçadaki örnek sayısının diğer k-1 parçadaki örnek sayılarından daha az olacağı unutulmamalıdır). Bu durumda k adet analiz ardışık olarak gerçekleştirilmektedir. Çapraz geçerleme sürecinde sırasıyla k parçanın her biri test veri seti olarak kullanılırken diğer k-1 parçalar eğitim veri seti olarak kullanılmaktadır (Bramer, 2013). Örnek olması bakımından k=5 için elde edilen sonuçların değerlendirilmesi amacıyla kullanılacak

olan 5 katlı çapraz geçerleme sürecinde nasıl bir yol izlendiği Şekil 2’de görsel olarak verilmiştir.



Şekil 2. Çapraz geçerlik süreci

Çapraz geçerlemede sabit bir katlanma ya da verilerin bölünme sayısına karar verilir. Örneğin şekil 2’de görüldüğü gibi kullanılan sabit sayı 5 olsun. Daha sonra veriler yaklaşık olarak eşit beş parçaya ayrılır ve her biri sırasıyla test için kullanılırken geri kalanlar eğitim için kullanılır. Yani test için verilerin beşte biri kullanılırken eğitim için eldeki verilerin beşte dördü kullanılmış olur. Böylece her bir örnek test için bir kez kullanılmış olacaktır. Buna beş katlı çapraz geçerlik denir ve eğer tabakalandırma işlemi yapılmışsa buna tabakalandırılmış beş katlı çapraz geçerlik denilmektedir. Bu yöntemi kullanmak eğitim ve test verilerindeki dengesiz bir temsili ortadan kaldırmak için ilkel bir koruma olarak görülmektedir (Vanwinckelen ve Blockeel, 2012).

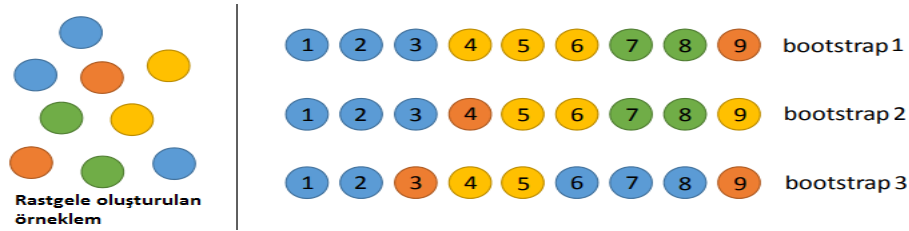
Elde tek ve sabit bir veri olduğunda öğrenme yönteminin hata oranını tahmin edebilmek için standart olarak 10 katlı çapraz geçerleme yaklaşımı uygulanmaktadır. Bu yaklaşımda verilerin tamamının yer aldığı çalışma evreni rastgele 10 sınıfa ayrılmaktadır ve her bir sınıfın tam veri kümesinde yaklaşık olarak aynı oranda temsil edilmesi gerekmektedir. Her bir sınıf sırasıyla test için ayrı bir kenarda bekletilirken geriye kalan dokuz sınıf üzerinden öğrenme süreci gerçekleşir ve bir kenarda bekletilen 1/10’luk veriler üzerinden hata oranı hesaplanmaktadır. Böylece öğrenme, farklı eğitim verileri üzerinden 10 kez tekrar edilerek gerçekleşmektedir. Son olarak 10 farklı hata teriminin ortalaması alınarak genel bir hata terimi elde edilmektedir (Witten ve Frank, 2005).

On katlı çapraz geçerleme yöntemi uygulamada herkes tarafından kullanılan standart bir değerlendirme ölçütü durumuna gelmiştir. Deneysel çalışmalarda tabakalı örnekleme yaklaşımının sonuçları biraz daha iyileştirdiği görülmektedir (Vanwinckelen ve Blockeel, 2012). Her ne kadar performans tahmininin doğru ve tutarlı bir şekilde yapılması yöntemin bir üstünlüğü olarak görülse de birbiriyle çakışan eğitim verileri, karşılaştırma için 1. tip hatanın artması, karşılaştırma için performans varyansının tahmin edilememesi ve serbestlik derecesinin aşırı tahmin edilmesi gibi sınırlılıkları da bulunmaktadır (Refaeilzadeh, Tang ve Liu, 2009).

Bootstrap Yöntemi

Bu yöntem bir kestiricinin istatistiksel doğruluğunu değerlendirmek için Efron (1979) tarafından geliştirilmiştir. Bootstrap yöntemi örneklemin yer değiştirmesi olarak tanımlanabilecek istatistiksel tekniklere dayanmaktadır. Bu yöntemde eğitim veya test için eldeki veri setinden bir örneklem alınır ve seçilen örneklerin üzeri çizilir (Kuonen, 2018). Yani aynı örnek bir kez seçildiyse tekrar seçilmez (Dekking, Kraaikamp, Lopuhaa ve Meester, 2005). Bu yaklaşım futbol takımı oluşturmaya benzer. Nasıl ki bir oyuncu farklı iki takımda yer alamıyorsa Bootstrap yönteminde de aynı örnek iki kez seçilemez. Çoğu öğrenme yöntemi aynı örneği iki kez kullanmaktadır. Eğer bir örnek eğitim kümesinde iki kez yer almışsa sonuçlarda farklılık oluşacaktır. Matematiksel olarak konuşmak gerekirse aynı nesnenin birden fazla görüntüsünün olması durumunda "küme" olarak tanımlanan gruplardan söz etmek anlamlı olmayacaktır.

Geleneksel yaklaşımın modern bir seçeneği olan Bootstrap yöntemi, daha gerçekçi modellerin uygulanmasına izin veren bilgisayar yoğunluklu bir yeniden örnekleme yöntemidir (Boss, 2003). Görsel açıdan Bootstrap yönteminin daha iyi anlaşılması için Şekil 3'ün incelenmesi yararlıdır. Şekilde görüldüğü üzere rastgele oluşturulmuş olan örnekleme yer alan birimler her bir adımda yerleri değiştirilerek yeniden örnekleme dahil edilmektedir.



Şekil 3. Bootstrap yönteminde örnekleme süreci.

Bootstrap (Tekrarlamalı Örnekleme) yöntemi temel olarak 4 adımda özetlenmektedir. Birinci adımda tüm veri setindeki örneklerin yerleri değiştirilerek rastgele bir örneklem oluşturulur. İkinci adımda örneklemden bazı istatistikler hesaplanır (Örneğin, örneklemin medyan değeri hesaplanır). Üçüncü adımda ilk iki adımdaki işlemler Bootstrap dağılımını elde edebilmek için N_b defa tekrarlanır. Dördüncü ve son adımda Bootstrap dağılımı kullanılarak güven aralığı ve standart hata gibi değerler hesaplanır (Galdi ve Tagliaferri, 2017).

Evet-hayır, ödünç verme-vermeme, kredi verme -vermeme, kuşku - kuşku değil vb. şekillerde iki sınıfı olan durumlarda olası sonuçlar Tablo 1'de gösterildiği gibidir. Doğru pozitif (DP) ve doğru negatif (DN) doğru sınıflama sonuçlarıdır. DP doğru tahmin edilen örnek sayısıdır. DN doğru olarak reddedilen örneklerin sayısıdır.

Yanlış pozitif (YP), sonucun aslında negatif iken yanlışlıkla pozitif; yanlış negatif (YN), sonucun aslında pozitif iken yanlışlıkla negatif olarak tahmin edilmesidir (Williams, 2011). Tablo 1’de iki farklı sonucu olan bir tahmine ilişkin gerçek ve test sonuçlarının alabileceği durumlar gösterilmektedir (Schwenke ve Schering, 2007).

Tablo 1.

İki Sınıflı Bir Tahmine İlişkin Farklı Sonuçlar

		Gerçek Durum		Toplam
		Pozitif	Negatif	
Test Sonucu	Pozitif	Doğru Pozitif (DP)	Yanlış Pozitif (YP)	DP+YP
	Negatif	Yanlış Negatif (YN)	Doğru Negatif (DN)	YN+DN
	Toplam	DP+YN	YP+DN	DP+YP+YN+DN

İki sınıflı bir tahmin süreci sonucunda DP, YP, YN ve DN oranları aşağıda gösterildiği şekilde elde edilmektedir (Elhamahmy, Elmahdy ve Saroit, 2010).

$$\text{Yanlış pozitif oranı (YPO): } YP / (YP+DN)$$

$$\text{Doğru pozitif oranı (DPO): } DP / (DP+YN)$$

$$\text{Yanlış negatif oranı (YNO): } YN / (YN+DP)$$

$$\text{Doğru negatif oranı (DNO): } DN / (DN+YP)$$

Sonuçlara ilişkin oranların ardından genel başarı oranı doğru sınıflamaların sayısının toplam sınıflama sayısına bölümüdür.

$$\text{Genel Başarı} = \frac{DP+DN}{DP+DN+YP+YN} \quad (1)$$

Bu sınıflama için hata oranı ise 1’den “genel başarı” oranının çıkarılması sonucunda elde edilmektedir. Duyarlılık ve seçicilik kavramlarının da bilinmesinde yarar vardır.

Duyarlılık, testin gerçek pozitif durumlar içinden pozitif olan durumları ayırma yeteneğidir.

$$\text{Duyarlılık (Sensitivity)} = \frac{DP}{DP+YN} \quad (2)$$

Seçicilik, testin gerçek negatif durumlar içinden negatif olan durumları ayırma yeteneğidir.

$$\text{Seçicilik (Specifity)} = \frac{DN}{DN+YP} \quad (3)$$

Duyarlılık ve seçicilik değerleri, testin araştırılan durumla ilgili olan ve olmayanları birbirinden ne kadar iyi ayırt ettiğini tanımlamaktadır. Çok sınıflı bir tahminde testten elde edilen sonuçlar, karışıklık (confusion) matrisi adı verilen iki boyutlu, her bir sınıf için satır ve sütunları olan bir tablo üzerinde gösterilmektedir. Matrisin her bir elemanı satırı gerçek değeri, sütunu ise tahmin edilen (test sonucu) değeri gösterecek şekilde tanımlanmaktadır (Bazı durumlarda satırlar test sonuçlarını, sütunlar ise gerçek durumu gösterebilir.). İyi sonuçlar esas köşegen üzerindeki sayıların oldukça büyük ve köşegen dışındaki sayıların küçük (ideali sıfır) olması durumunda elde edilir (Elhamahmy, Elmahdy ve Saroit, 2010).

Araştırmanın Amacı

Veri madenciliğinde kuramsal temelleri farklı istatistiksel modellere dayalı olan birçok farklı öğrenme yöntemi bulunmaktadır. Bu yöntemlerin fazla olması nedeniyle hangilerinin daha etkili kestirimlerde bulunduğu ve hangilerinin daha az hatalı hesaplamalar yaptığının belirlenmesi, sonuçların güvenilirliği ve geçerliği bakımından önem taşımaktadır (Souza, Matwin ve Japkowicz, 2002). Model değerlendirmesinde kullanılan öğrenme yöntemleri arasında oluşacak belirgin farklılıkların tesadüfü etkilerden kaynaklanmadığından emin olmak için istatistiksel karşılaştırmalara gereksinim duyulmaktadır (Chamatkar ve Butey, 2014). Bu nedenle çalışma kapsamında veri madenciliğinde kullanılan farklı öğrenme yöntemleriyle elde edilen sonuçların güvenilirlik ve geçerlik değerlerinin farklı koşullar altında nasıl bir değişkenlik gösterdiğinin belirlenmesi amaçlanmıştır. Bu amaç kapsamında “PISA sınavına katılan öğrencilerinin Fen okuryazarlığı bakımından başarılarını tahmin etmede kullanılan veri madenciliği sınıflama algoritmalarının güvenilirlik ve geçerlik değerleri ne düzeydedir?” sorusuna cevap aranmıştır.

Araştırmanın alt problemleri aşağıdaki gibi belirlenmiştir:

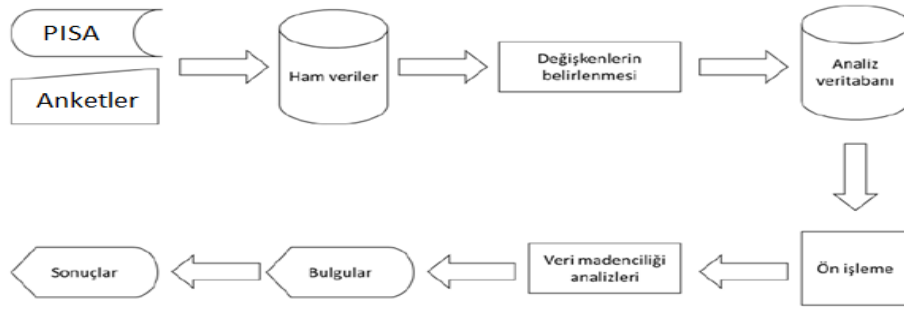
1. PISA öğrenci anketiyle ölçülen değişkenlerden yararlanarak “Fen başarı”nı tahminde kullanılan Decision Stump, Hoeffding Tree, J.48, Lojistik Model, RepTree, Random Forest, Random Tree ve Ridge lojistik Regresyon yöntemleriyle edilen sonuçların, 10 katlı çapraz geçirme yöntemi yardımıyla elde edilen güvenilirlik ve geçerlik değerleri ne düzeydedir?
2. PISA öğrenci anketiyle ölçülen değişkenlerden yararlanarak “Fen başarı”nı tahminde kullanılan Decision Stump, Hoeffding Tree, J.48, Lojistik Model, RepTree, Random Forest, Random Tree ve Ridge lojistik Regresyon yöntemleri yardımıyla edilen sonuçların, test edilen verilerin %5, %10, %15, %20, %25, %30 ve %35 olması durumunda güvenilirlik ve geçerlik değerleri ne düzeydedir?
3. PISA öğrenci anketiyle ölçülen değişkenlerden yararlanarak “Fen başarı”nı tahminde kullanılan Decision Stump, Hoeffding Tree, J.48, Lojistik Model, RepTree, Random Forest, Random Tree ve Ridge lojistik Regresyon yöntemleri yardımıyla edilen sonuçların, karışıklık matrisi yardımıyla elde edilen doğruluk dereceleri ne düzeydedir?

Yöntem

Bu bölümde araştırmanın türü, çalışma grubu, araştırma koşulları ve verilerin analizine ilişkin bilgilere yer verilmektedir.

Araştırma Modeli

Araştırmada izlenen süreç ve sürece ilişkin her bir bileşen görsel olarak daha iyi anlaşılması amacıyla Şekil 4'te gösterilmiştir.



Şekil 4. Araştırma süreci.

Öğrencilerinin PISA öğrenci anketinde yer alan alt ölçeklere verdikleri yanıtlar yardımıyla Fen okuryazarlığı başarılarını tahmin etmek ve bu süreçte kullanılan veri madenciliği tahminleme yöntemlerinin güvenilirlik ve geçerlik değerlerinin ne düzeyde olduğunu belirlemek amacıyla Şekil 4'te verilen adımlar izlenmiştir. Buna göre öncelikle PISA öğrenci anketinden ham veriler elde edilmiş ve ardından değişkenler belirlenmiştir. Ardından analiz edilecek veriler ön işleme tabi tutulmuş ve veri madenciliği yöntemleri ile analiz edilmiştir. Analiz sonucunda elde edilen bulgular ve sonuçlara dayalı olarak önerilerde bulunulmuştur.

Araştırmada PISA fen başarısını tahmin etmede kullanılan farklı veri madenciliği algoritmalarına ilişkin güvenilirlik ve geçerlik değerlerinin ne düzeyde olduğunu belirlemek amacıyla çalışmada amaçlandığınan çalışma betimsel araştırma niteliğindedir (Fraenkel, Wallen ve Hyun, 2012).

Evren-Örneklem

Araştırmanın amaçları doğrultusunda çalışma evreni OECD tarafından düzenlenen PISA 2015 sınavına katılan ve örgün eğitime kayıtlı olan 15 yaş grubu öğrencileri oluşturmaktadır. Sınava 72 ülkeden toplam 540.000 civarında öğrenci katılmış ve bunların 5895'i Türkiye'den katılan öğrencilerdir. PISA 2015 Türkiye uygulamasında 15 yaş grubu öğrenci evreni 1.324.089 öğrenci, uygulamaya katılabilecek ulaşılabilir Türkiye evreni ise 925.366 öğrenci olarak belirlenmiştir. PISA araştırmasında örneklem, tabakalı seçkisiz yöntemle belirlenmektedir (MEB, 2016).

Veri Toplama Süreci

Araştırmada kullanılan veriler 2017 yılında paylaşıma açılan ve OECD'nin resmi internet sayfası olan <http://www.oecd.org/pisa/data/2015database/> adresinden elde edilmiştir. SPSS veri dosyası formatında yer alan ve ülke kodu TR olan 5895 öğrenciye ilişkin veriler analize dahil edilmiştir.

Veri Toplama Araçları

Araştırmada veri toplama aracı olarak kullanılan ve PISA öğrenci anketinde yer alan, aşağıda isimleri ve kodları yazılı olan değişkenler Tablo 2'de gösterilmiştir.

Tablo 2

Çalışma Kapsamında Kullanılan Değişkenlerin Adları ve Kodları

Değişkenin Adı	Kodu	Değişkenin Adı	Kodu
Disiplin ortamı	DISCLISCI	Fen öğrenme süresi (hafta)	SMINS
Öğretmen desteği	TEACHSUP	Toplam öğrenme süresi (hafta)	TMINS
Sorgulamaya dayalı fen eğitimi	IBTEACH	Okula ait hissetme	BELONG
Öğretmen merkezli eğitim	TDTEACH	Test kaygısı	ANXTEST
Çevreye duyarlılık	ENVAWARE	Tutum, tercih ve öz inançlar	MOTIVAT
Fenden keyif alma	JOYSCIE	Birlikte çalışmaya isteği	COOPERATE
Fen konularına ilgi	INTBRSCI	Ailenin duygusal desteği	EMOSUPS
Araçsal motivasyon	INSTSCIE	Algılanan geri bildirim	PERFEED
Fen özyeterliliği	SCIEEFF	Öğretmenin adil olması	unfairteacher
Bilimsel inançlar	EPIST	Evdeki eğitim kaynakları	HEDRES
Fen yaşantısı	SCIEACT	Evdeki eğitimsel eşyalar	HOMEPOS
Öğrenciden beklenen mesleki statü	BSMJ	BİT kaynakları	ICTRES
Anne eğitim düzeyi	MISCED	Ailenin mal varlığı	WEALTH
Baba eğitim düzeyi	FISCED	Sosyo ekonomik durum indeksi	ESCS
Okul dışı ders çalışma süresi (hafta)	OUTHOURS	Fen Okuryazarlığı	PV1SCIE

Tablo 2'de bağımsız değişkenler ile fen okuryazarlığı bağımlı değişkeninin veri dosyasındaki kodları görülmektedir. PISA sınavında kullanılan değişkenler indeks değişkenleridir. Örneğin her bir öğrenci için hesaplanan sosyo-ekonomik ve kültürel durum indeksi (EKSD), anne-babanın mesleği ve eğitim düzeyi, öğrencinin evde sahip olduğu eğitim ile ilgili kaynaklar ve ailenin evindeki bir takım araç gereçler değişkenleri dikkate alınarak hesaplanmaktadır (MEB, 2015). PISA uygulamalarında sorular/maddeler 4'lü Likert tipi derecelendirilmiş şekilde yapılandırılmıştır. Bu sorular/maddeler düzeyinde toplam puanlar elde edilip ortalama ve standart sapmalarına göre standartlaştırılarak (-5, +5) aralığında indeks puanlar oluşturulmaktadır (Tatlı, Ergin ve Demir, 2016). Her bir duyuşsal özelliğe yönelik

ölçme sonuçları bu indekslerle tanımlanmaktadır. Yüksek indeks değeri, ilgili özelliğin nicelik olarak daha büyük olduğu anlamına gelmektedir. Araştırmada bilgi teknolojileri kullanımı, öğrencinin geçmiş fen deneyimi, ailenin öğrenme için sağladığı destek, öğrencinin değiştirdiği okul sayısı ve ailenin algıladığı okul kalitesi değişkenleri Türkiye örnekleminde boş bırakılması ve/veya cevaplanmaması nedeniyle analize dahil edilememiştir.

Verilerin Analizi

Araştırmanın ilk aşamasında öğrencilerin PISA Fen okuryazarlık puanları Türkiye ortalaması olan 425,00 puanının altında olanlar başarısız (0), bu puanın üzerinde olanlar başarılı (1) olarak kodlanmıştır. Elde edilen iki kategorili sonuçlara göre 5865 öğrencinin 2769'u başarılı (%47,20), 3096'sı başarısız (%52,8) olarak sınıflanmıştır. Araştırmanın ikinci aşamasında bağımsız değişkenlerden yararlanarak öğrencilerin PISA fen okuryazarlığı performanslarını tahmin edecek bir model oluşturulmuştur (Hastie, Tibshirani ve Friedman, 2009). Öğrencilerin PISA sınavında gösterdikleri performans “Başarılı” ve “Başarısız” şeklinde kodlandığı için bu bir sınıflama problemi ve veri madenciliği sınıflama yöntemleri kullanılarak tahmin yapılmıştır (Olmo, Romero ve Ventura, 2012). Bu amaçla alanyazında sınıflama ve tahmin amacıyla sıklıkla kullanılan Decision Stump, Hoeffding Tree, J.48, Lojistik Model, RepTree, Random Forest, Random Tree ve Ridge lojistik Regresyon yöntemlerinden yararlanılmıştır. Araştırmanın üçüncü aşamasında, araştırma kapsamında belirlenen sınıflama algoritmalarıyla başarılı ve başarısız olarak sınıflanan öğrenciler için güvenilirlik ve geçerlik değerleri incelenmiştir.

Çalışmanın alt problemlerine geçmeden önce fen okuryazarlığını yordamak amacıyla kullanılan değişken sayısının PISA öğrenci anketi esas alındığında çok fazla olması sebebiyle öncelikle çalışma kapsamında kullanılacak değişkenler belirlenmiştir. Her ne kadar YSA yöntemleri regresyon analizi yöntemine göre modele eklenen değişken sayısının fazla olması durumunda daha iyi tahmin yapacağı belirtilse de (Lykourentzou, Giannoukos, Mpardis, Nikolopoulos ve Loumos, 2009) VM gibi tahmine dayalı yöntemlerde modele daha çok sayıda değişken eklemenin performans kestiriminin doğruluğunda bir artışa neden olmayacağını belirtmiştir (Huang ve Fang, 2013). Bunun yanında Kohavi (1995) tarafından yapılan deneysel çalışmada farklı algoritmalar altında yapılan kestirimlerin katman sayısının 10 ve 20 olması durumunda oldukça iyi olduğu ve neredeyse tamamen yansız olduğu belirtildiği için çalışmada 10 katlı çapraz geçişleme yöntemi kullanılmıştır. Değişken sayısını azaltmak amacıyla öncelikle VM yöntemlerinden Best First Forward, Best First Backward ve Greedy Stepwise yardımıyla 10 katlı çapraz geçişleme yöntemi sonucunda en az %40 ve üzeri başarılı olan toplam 12 değişken yardımıyla PISA fen okuryazarlığı tahmin edilmeye çalışılmıştır.

Çalışmada güvenilirlik ölçütleri olarak doğru sınıflama sayısı, doğru sınıflama oranı, Kappa istatistiği, ortalama mutlak hata, karekök hata, göreceli mutlak hata ve göreceli karekök hata değerleri kullanılırken ve geçerlik ölçütleri olarak doğru pozitif oranı, yanlış pozitif oranı, duyarlılık, geri çağırma, F-ölçütü, Matthew korelasyon

katsayısı, ROC eğrisi altında kalan alan ve Duyarlık-Geri çağırma eğrisi altında kalan alan değerleri kullanılmıştır. Özellikle kesme puanı belirleme aşamasında duyarlık (sensitivity), özgüllük (specificity), genel doğruluk (accuracy) ve eğri altında kalan alan (ROC) gibi yöntemlerle elde edilen kesme puanının etkinliği ve verimliliği hakkında yorum yapılabilmektedir (Kumar ve Indrayan, 2011). ROC eğrisi, tanı koymak amacıyla kullanılan bir değişkenin değişim aralığı içinde değişkenin tüm değerleri sırasıyla kesim noktası kabul edilmesiyle hesaplanacak duyarlılık değerlerinin, testin yanlış pozitif oranına (1 - özgüllük) karşı noktalanması ile elde edilir. ROC eğrisinin oluşturulacağı koordinat sisteminde, Y ekseninde tanı testinin gerçek pozitif değeri (duyarlılık), X ekseninde ise yanlış pozitif değeri (1-özgüllük) yer alır (Egan, 1975). ROC eğrisi altındaki alanın beklenen değeri 0.50'dir. Mükemmel bir test ise sıfır yanlış pozitif ve sıfır yanlış negatif ile alanın değeri 1.00 olacaktır. Eğri altında kalan alanın yorumlanmasında; 90-1.00 = mükemmel, .80-.90 = iyi, .70-.80 = orta, .60-.70 = zayıf ve .50-.60 = başarısız olarak sınıflandırılmaktadır.

Bulgular

Çalışma kapsamında ilgili alanyazın taraması sonucunda öğrencilerin fen okuryazarlığı üzerinde etkili olduğu belirlenen değişkenlerin isimleri, kodları, en düşük ve en yüksek değerleri ile aritmetik ortalamaları Tablo 3'te gösterilmiştir.

Tablo 3

Çalışma Kapsamında Kullanılacak Değişkenlere İlişkin İstatistikler

Değişkenin Adı	Kodu	Min.	Max.	Ort.	Kayıp Oranı (%)
Disiplin ortamı	DISCLISCI	-2.416	1.883	-0.134	10.7
Öğretmen desteği	TEACHSUP	-2.719	1.447	0.196	10.9
Sorgulamaya dayalı fen eğitimi	IBTEACH	-3.340	3.182	0.320	10.9
Öğretmen merkezli eğitim	TDTEACH	-2.447	2.078	-0.059	11.1
Çevreye duyarlık	ENVAWARE	-3.376	3.293	0.552	4.2
Fenden keyif alma	JOYSCIE	-2.115	2.163	0.124	4.4
Fen konularına ilgi	INTBRSCI	-2.581	2.730	-0.065	8.5
Araçsal güdüleme	INSTSCIE	-1.930	1.735	0.375	5.1
Fen özyeterliği	SCIEEFF	-3.756	3.277	0.336	5.1
Bilimsel inançlar	EPIST	-2.790	2.155	-0.192	4.8
Fen yaşantısı	SCIEACT	-1.757	3.361	0.687	5.3
Öğrenciden beklenen mesleki statü	BSMJ	10	89	63.35	7.7
Anne eğitim düzeyi	MISCED	0	6	2.20	1.6
Baba eğitim düzeyi	FISCED	0	6	2.67	1.6
Okul dışı ders çalışma süresi (hafta)	OUTHOURS	0	70	25.54	8.8
Fen öğrenme süresi (hafta)	SMINS	0	800	198.82	4.7

(devam ediyor)

Tablo 3 (devam)

Değişkenin Adı	Kodu	Min.	Max.	Ort.	Kayıp Oranı (%)
Toplam öğrenme süresi (hafta)	TMINS	100	3000	1558.75	6.5
Okula ait hissetme	BELONG	-3.129	2.612	-0.436	1.5
Test kaygısı	ANXTEST	-2.505	2.549	0.318	1.3
Tutum, tercih ve öz inançlar	MOTIVAT	-3.087	1.854	0.613	1.5
Birlikte çalışmaya isteği	COOPERATE	-3.332	2.287	0.000	1.6
Ailenin duygusal desteği	EMOSUPS	-3.078	1.099	-0.267	1.1
Algılanan geri bildirim	PERFEED	-1.525	2.499	0.351	11.3
Öğretmenin adil olması	unfairteacher	1	24	10.25	1.8
Evdeki eğitim kaynakları	HEDRES	-4.370	1.180	-0.583	1.5
Evdeki eğitimsel eşyalar	HOMEPOS	-6.710	5.150	-1.432	0.6
BİT kaynakları	ICTRES	-3.270	3.500	-1.190	1.2
Ailenin mal varlığı	WEALTH	-6.960	4.090	-1.487	0.9
Sosyo ekonomik durum indeksi	ESCS	-5.130	3.120	-1.447	0.6
Fen okuryazarlığı	PV1SCIE	197.72	707.89	422.45	0

Tablo 3 incelendiğinde veri madenciliği yöntemleriyle fen okuryazarlığını yordamak amacıyla kullanılacak değişken sayısının 29 olduğu belirlenmiştir. Bu işlemlerin ardından kayıp veri içermeyen tam verilerin Weka programına aktarılması amacıyla gerekli dönüşümler yapılmış ve analizlere geçilmiştir.

PISA Fen okuryazarlığını yordamak amacıyla kullanılacak değişken sayısının çok fazla (29) olmasından dolayı farklı algoritmalar yardımıyla başarıyı en iyi yordayan özellikler belirlenmiştir. Bu amaçla eğitim veri seti üzerinden kullanılan algoritmalar BestFirst ve Greedy Stepwise olup her bir yöntemle elde edilen sıralama sonuçları Tablo 4'te gösterilmiştir.

Tablo 4

PISA Okuryazarlığını En İyi Yordayan Değişkenler

Sıra No	Best First -forward	Best First -backward	Greedy Stepwise
1	Sorgulamaya dayalı fen eğitimi	Sorgulamaya dayalı fen eğitimi	Sorgulamaya dayalı fen eğitimi
2	Çevreye duyarlık	Çevreye duyarlık	Çevreye duyarlık
3	Bilimsel inançlar	Bilimsel inançlar	Bilimsel inançlar
4	Öğrenciden beklenen mesleki statü	Öğrenciden beklenen mesleki statü	Öğrenciden beklenen mesleki statü
5	Okul dışı ders çalışma süresi	Okul dışı ders çalışma süresi	Okul dışı ders çalışma süresi
6	Fen öğrenme süresi	Fen öğrenme süresi	Fen öğrenme süresi
7	Toplam öğrenme süresi	Toplam öğrenme süresi	Toplam öğrenme süresi

(devam ediyor)

Tablo 4 (devam)

Sıra No	Best First -forward	Best First -backward	Greedy Stepwise
8	Öğretmenin adil olması	Öğretmenin adil olması	Öğretmenin adil olması
9	Evdeki eğitim kaynakları	Evdeki eğitim kaynakları	Evdeki eğitim kaynakları
10	BİT Kaynakları	BİT Kaynakları	BİT Kaynakları
11	Sosyo ekonomik durum indeksi	Sosyo ekonomik durum indeksi	Sosyo ekonomik durum indeksi

Tablo 4 incelendiğinde fen okuryazarlığını en iyi yordayan ilk 11 değişkenin farklı yöntemler kullanıldığında değişmediği belirlenmiştir. Elde edilen bu sonucun geçerliği 10 katlı çapraz geçişleme yöntemiyle test edilmiştir. On katlı çapraz geçişleme analizlerinde değişkenlerin en az 4 katmanda anlamlı etkiye sahip olması ölçüt alınmıştır. Modele dahil edilecek değişkenler, kodları ve başarılı olunan katman sayısı Tablo 5’te gösterilmiştir.

Tablo 5

Veri Madenciliğinde Kullanılacak Değişkenlere İlişkin Özet Bilgiler

Sıra No	Değişkenin Adı	Kodu	Min	Max	Başarı Oranı (%)
1	Sorgulamaya dayalı fen eğitimi	IBTEACH	-3.340	3.182	100
2	Çevreye duyarlık	ENVAWARE	-3.376	3.293	100
3	Bilimsel inançlar	EPIST	-2.790	2.155	100
4	Öğrenciden beklenen mesleki statü	BSMJ	10	89	100
5	Okul dışı ders çalışma süresi (hafta)	OUTHOURS	0	70	100
6	Fen öğrenme süresi (hafta)	SMINS	0	800	100
7	Toplam öğrenme süresi (hafta)	TMINS	100	3000	100
8	Test kaygısı	ANXTEST	-2.505	2.549	60
9	Öğretmenin adil olması	Unfairteacher	1	24	40
10	Evdeki eğitim kaynakları	HEDRES	-4.370	1.180	80
11	BİT kaynakları	ICTRES	-3.270	3.500	100
12	Sosyo ekonomik durum indeksi	ESCS	-5.130	3.120	100

Bu işlemlerin ardından 12 bağımsız değişken için PISA fen okuryazarlığını yordamak amacıyla kullanılan veri madenciliği yöntemlerinin güvenilirlik ve geçerlik değerlerinin karşılaştırılması aşamasına geçilmiştir.

Birinci Alt Probleme İlişkin Bulgular

PISA Fen başarısını tahmin etmede kullanılan Decision Stump, Hoeffding Tree, J.48, Lojistik Model, RepTree, Random Forest, Random Tree ve Ridge lojistik Regresyon yöntemleri yardımıyla edilen sonuçların 10 katlı çapraz geçişleme yöntemi yardımıyla elde edilen güvenilirlik ve geçerlik değerleri ne düzeydedir?

Farklı yöntemlerin karşılaştırılması. Farklı yöntemler yardımıyla elde edilen güvenilirlik değerlerinin karşılaştırılması amacıyla ağırlıklandırılmış genel ortalama değerleri karşılaştırılmıştır. Her bir yöntem için doğru sınıflanan örnek sayıları, doğru sınıflama oranları, Kappa istatistiği, mutlak hatanın ortalaması, hataların ortalama karekökü ve göreceli mutlak hata değerleri Tablo 6’da gösterilmiştir.

Tablo 6

Farklı Yöntemler Kullanılarak Elde Edilen Güvenirlik Ölçütleri

Algoritma	Doğru Sınıflama Sayısı	Doğru Sınıflama Oranı	Kappa İstatistiği	Ortalama Mutlak Hata	Karekök Hata	Görelî Mutlak Hata	Görelî Karekök Hata
1. Decision Stump	3744	63.83	0.276	0.460	0.479	92.32	96.10
2. Hoefding Tree	4085	69.65	0.390	0.357	0.455	71.76	91.18
3. J.48	3976	67.79	0.351	0.382	0.492	76.79	98.67
4. Lojistik Model	4125	70.33	0.403	0.381	0.437	76.58	87.70
5. RepTree	3946	67.28	0.341	0.391	0.472	78.56	94.65
6. Random Forest	4177	71.22	0.420	0.378	0.434	76.01	86.92
7. Random Tree	3690	62.92	0.256	0.370	0.609	74.40	121.98
8. Ridge Lojistik Reg.	4141	70.61	0.409	0.384	0.438	77.11	87.82

Tablo 6 incelendiğinde doğru sınıflama sayısı, doğru sınıflama oranı, kappa istatistiği, karekök hata ve göreceli karekök hata değerleri bakımından en iyi sonuçların Random Forest yöntemiyle elde edilirken mutlak hatanın ortalaması ve göreceli mutlak hata bakımından en iyi sonuçların Hoefding Tree yöntemiyle elde edildiği belirlenmiştir. Bunun yanında doğru sınıflanan örnek sayısı ve oranı, Kappa istatistiği, hataların ortalama karekökü ve göreceli hataların karekökü bakımından en kötü sonuçların Random tree tarafından elde edilirken mutlak hatanın ortalaması ve göreceli mutlak hata bakımından en kötü sonuçların Decision Stump yöntemiyle elde edildiği belirlenmiştir.

Bu araştırmada Decision stump, Hoefding tree, J.48, Lojistik model, RepTree, Random Forest, Random Tree ve Ridge Lojistik Regresyon yöntemleri kullanılarak başarılı ve başarısız olarak sınıflama sonuçları için elde edilen ağırlıklandırılmış ortalamalara dayalı geçerlik ölçütleri Tablo 7’de gösterilmiştir.

Tablo 7

Farklı Yöntemler Kullanılarak Elde Edilen Geçerlik Ölçütleri

Algoritma	DP Oranı	YP Oranı	Duyarlık	Geri Çağırma	F-Ölçüt	Matthew	ROC Alanı	DG Eğrisi
1. Decision Stump	0.638	0.361	0.640	0.638	0.639	0.277	0.629	0.596
2. Hoefding Tree	0.697	0.307	0.696	0.697	0.696	0.390	0.758	0.739
3. J.48	0.678	0.328	0.678	0.678	0.677	0.352	0.694	0.649
4. Lojistik Model	0.703	0.301	0.703	0.703	0.703	0.404	0.778	0.767
5. RepTree	0.673	0.332	0.672	0.673	0.672	0.342	0.718	0.688
6. Random Forest	0.712	0.294	0.712	0.712	0.711	0.421	0.785	0.777
7. Random Tree	0.629	0.373	0.629	0.629	0.629	0.256	0.628	0.582
8. Ridge Lojistik	0.706	0.298	0.706	0.706	0.706	0.409	0.777	0.765

Tablo 7 incelendiğinde Random Forest yöntemiyle elde edilen tüm geçerlik ölçütlerinin diğer yöntemlerden daha yüksek olduğu belirlenmiştir. ROC eğrisi altında kalan alanlar karşılaştırıldığında Decision Stump, J.48 ve Random tree algoritmaları ile elde edilen sonuçların zayıf ve diğer algoritmalar tarafından elde edilen sonuçların orta düzeyde olduğu belirlenmiştir. Güvenirlik ve geçerliğe ilişkin elde edilen sonuçlar bir bütün olarak değerlendirildiğinde eğer kategorik bir bağımlı değişken için bireyleri sınıflamak istenirse en iyi sonuçların sırasıyla Random forest, Ridge lojistik regresyon, Lojistik model, Hoefding tree, J.48, RepTree, Decision Stump ve Random Tree algoritmaları şeklinde olacağı görülmektedir. Eğer Kappa istatistiği dikkate alınarak bir sıralama yapmak istenirse en iyi sonuçların sırasıyla Random forest, Ridge lojistik regresyon, Lojistik model, Hoefding tree, J.48, RepTree, Decision Stump ve Random Tree algoritmaları şeklinde olacağı görülmektedir. Buna göre doğru sınıflama veya kappa istatistiğini esas almanın sıralamada bir değişikliğe neden olmayacağı görülmektedir.

İkinci Alt Probleme İlişkin Bulgular

PISA Fen başarısını tahmin etmede kullanılan Decision Stump, Hoefding Tree, J.48, Lojistik Model, RepTree, Random Forest, Random Tree ve Ridge Lojistik Regresyon yöntemleri yardımıyla edilen sonuçların test edilen verilerin %5, %10, %15, %20, %25, %30 ve %35 olması durumunda güvenirlik ve geçerlik sonuçları nasıldır?

Farklı yöntemlerden elde edilen sonuçların karşılaştırılması. Genel olarak bakıldığında test edilen veri setinin %15 olması durumunda değerlendirme ölçütlerinin farklılaştığı ve ölçütlerin tutarlı sonuçlar üretmediği görülmektedir. Öğrenme yöntemleri hata ölçütleri bakımından kendi içinde karşılaştırıldığında en tutarlı sonuçlar Ridge regresyon yöntemi ve Lojistik model tarafından elde edilmektedir. Elde edilen sonuçlardan doğru pozitif oranlarının farklı yüzdeliklerdeki test verilerinde gösterdiği değişimler Tablo 8’de gösterilmiştir.

Tablo 8

Farklı Yöntemler Kullanılarak Elde Edilen Doğru Pozitif Oranları

DP Oranı	%5	%10	%15	%20	%25	%30	%35
1. Decision Stump	71.31	67.91	64.43	64.27	63.50	63.44	63.22
2. Hoefding Tree	70.64	69.96	70.22	69.99	67.80	66.00	63.17
3. J.48	69.96	68.08	66.02	68.28	67.80	65.88	67.65
4. Lojistik Model	71.67	71.84	73.06	71.95	71.14	70.60	70.53
5. RepTree	66.89	67.57	71.25	67.17	68.62	66.62	66.43
6. Random Forest	72.01	71.16	73.75	72.37	73.46	70.83	69.99
7. Random Tree	62.79	62.79	63.97	61.63	64.39	60.26	64.39
8. Ridge Lojistik R.	72.35	71.67	72.72	71.61	71.35	70.66	70.19

Tablo 8 incelendiğinde Lojistik model, Random Forest ve Ridge Regresyon yöntemlerinin farklı büyüklükteki test verilerinde en doğru tahmin sonuçlarını verirken test edilen veri seti oranı %15 olduğunda en yüksek değere ulaştığı sonra tekrardan azaldığı görülmektedir. Hataların ortalama karekökünün farklı test veri setlerinde ve farklı yöntemlerde nasıl bir değişim gösterdiği Tablo 9’da verilmiştir.

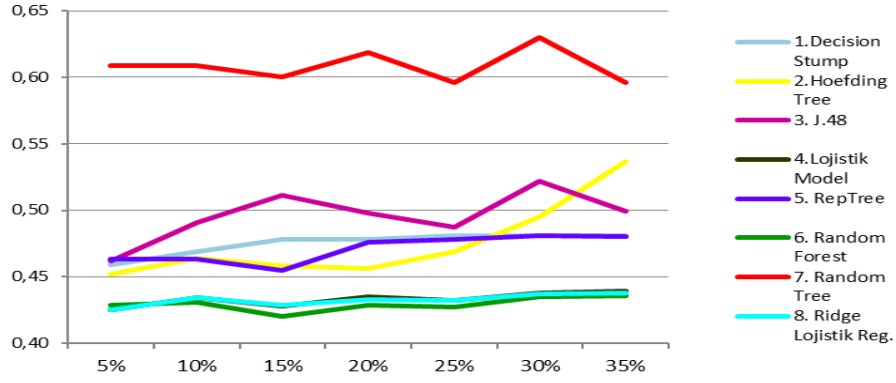
Tablo 9

Farklı Yöntemler Kullanılarak Elde Edilen Hataların Ortalama Karekökü

Hata	%5	%10	%15	%20	%25	%30	%35
1. Decision Stump	0.459	0.469	0.478	0.478	0.481	0.481	0.481
2. Hoefding Tree	0.452	0.464	0.458	0.456	0.469	0.495	0.537
3. J.48	0.462	0.491	0.511	0.498	0.487	0.522	0.499
4. Lojistik Model	0.425	0.434	0.428	0.435	0.432	0.438	0.439
5. RepTree	0.463	0.463	0.455	0.476	0.478	0.481	0.480
6. Random Forest	0.429	0.431	0.420	0.429	0.427	0.435	0.436
7. Random Tree	0.609	0.609	0.600	0.619	0.596	0.630	0.596
8. Ridge Lojistik R.	0.425	0.434	0.429	0.433	0.432	0.437	0.438

Tablo 9 incelendiğinde Lojistik model, Random Forest ve Ridge Regresyon yöntemleri farklı büyüklükteki test verilerinde en düşük hata değerlerini vermiştir. En düşük hata değerlerinin test edilen veri seti oranı %15 olduğunda elde edildiği ve daha sonra tekrar arttığı görülmektedir. Sonuçların daha iyi anlaşılması bakımından

Hataların Ortalama Karekökü değerlerinin farklı yöntemler altındaki değişimi Şekil 5'te gösterilmiştir.



Şekil 5. Farklı algoritmalar kullanılarak elde edilen hataların değişimi.

Şekil 5 incelendiğinde Random Tree yönteminin en yüksek hata değerine sahip olduğu ve ikinci en yüksek hata değerlerinin J.48 yöntemi ile elde edildiği görülmektedir. Bunun yanında en düşük hata değerlerinin Ridge regresyon, Random forest ve Lojistik modele ait olduğu görülmektedir. Ayrıca bu yöntemlerle elde edilen hata değerlerinin farklı yüzdelikteki test verilerinde de tutarlı olduğu belirlenmiştir.

Üçüncü Alt Probleme İlişkin Bulgular

PISA Fen başarısını tahmin kullanılan Decision Stump, Hoeffding Tree, J.48, Lojistik Model, RepTree, Random Forest, Random Tree ve Ridge Lojistik Regresyon yöntemleriyle edilen sonuçların karışıklık matrisiyle hesaplanan doğruluk dereceleri ne düzeydedir?

Çalışmanın üçüncü alt probleminde veri setini test ve eğitim verisi olarak ayırmayıp aynı veri seti üzerinden hem öğrenme yöntemini eğitip hem de test ettiğimizde elde edilen sonuçların doğruluk dereceleri belirlenmiştir. Aynı veri seti üzerinden hem eğitilip hem de test edilen öğrenme yöntemlerinin doğruluk oranları ile başarılı ve başarısız olarak tahmin ettiği öğrenci sayıları Tablo 10'da gösterilmiştir.

Tablo 10
Farklı Öğrenme Yöntemlerine İlişkin Karışıklık Matrisi ve Doğruluk Oranları

Öğrenme Yöntemi		Karışıklık Matrisi		Genel Başarı (Doğru Sınıflama Oran %)
		Başarılı	Başarısız	
Decision Stump	Başarılı	1959	1137	63.87
	Başarısız	982	1787	
Hoefding Tree	Başarılı	2296	800	68.88
	Başarısız	1025	1744	
J.48	Başarılı	2681	415	80.95
	Başarısız	702	2067	
Lojistik Model	Başarılı	2352	744	72.07
	Başarısız	894	1875	
RepTree	Başarılı	2578	518	77.59
	Başarısız	796	1973	
Random Forest	Başarılı	3096	0	100
	Başarısız	0	2769	
Random Tree	Başarılı	3096	0	100
	Başarısız	0	2769	
Ridge Regresyon	Başarılı	2290	806	70.74
	Başarısız	910	1859	

Tablo 10 incelendiğinde aynı veri seti üzerinde hem öğrenme yönteminin eğitilmesi hem de test edilmesi sonucunda Random Forest ve Random Tree yöntemlerinin tüm öğrencileri doğru sınıflayarak %100 doğru sınıflama oranına sahip oldukları belirlenmiştir. Bu yöntemlerin ardından en iyi doğru sınıflama oranına sahip öğrenme yönteminin J.48 algoritması olduğu ve başarı oranının % 80,95 olduğu belirlenmiştir. En düşük doğru sınıflama oranına sahip öğrenme yönteminin Decision Stump ve başarı oranının % 63,87 olduğu saptanmıştır. Her ne kadar 10 katlı çapraz geçişleme yöntemiyle başarı oranı %67,28 olan Random tree ve başarı oranı %67,79 olan J.48 yöntemleri, veri setinin eğitim ve test verisi olarak ayrılmaması durumunda sırasıyla %100 ve %80,95 gibi daha yüksek doğru sınıflama oranına sahip oldukları görülmektedir. Elde edilen bu sonucun aşırı uyum gösterme (over fitting) olarak bilinen sorundan kaynaklandığı düşünülmektedir. Zhou ve Cervantes (2016) aşırı uyum gösterme sorununun örneklemin evrenin iyi bir temsilcisi olmaması ve bu sebeple eğitim setindeki varyansın yüksek olması ile eğitim veri setindeki yanlış sınıf özellikleri ve yanlış sınıf etiketi gibi gürültülü durumun olmasından kaynaklandığını belirtmektedir. Bu sebeple yeterli delil olmaması durumunda ağaç büyümesinin durdurulması veya daha fazla düğümün oluşmasına engel olunması gibi çözüm yöntemleri önerilmektedir (Bramer, 2013). Araştırma sonucunda rastgele ağaç yönteminin yeterli kanıt olmamasına rağmen büyümeye devam etmesi sonucunda aslında gerçekçi olmayan %100 gibi bir başarı oranına sahip olduğu görülmektedir.

Tartışma, Sonuç ve Öneriler

Araştırmadan elde edilen sonuçlar alt problemlere ilişkin başlıklar altında açıklanmıştır.

Birinci Alt Probleme İlişkin Sonuçlar

Çalışmanın birinci alt probleminde verileri eğitim ve test verisi olarak ayırmadan doğrudan 10 katlı çapraz geçişleme yöntemi yardımıyla kullanılan veri madenciliği tahmin yöntemlerinin güvenilirlik ve geçerlik değerleri karşılaştırılmıştır. Kullanılan güvenilirlik ölçütlerinden doğru sınıflanan örnek sayısı, kapa istatistiği, mutlak hata ve ortalama hataların kareköküne göre en iyi tahmin yönteminin Random forest olduğu belirlenmiştir. Elde edilen bu sonuç Liaw ve Wiener (2002) ile Svetnik, Liaw, Tong ve Wang (2004) tarafından yapılan çalışmalarla benzerlik göstermektedir. Random forest yönteminde bir sınıflandırıcı yerine birden çok sınıflandırıcı kullanılması ve sonrasında onların tahminlerinden elde edilen sonuçlar ile yeni veriyi sınıflandıran öğrenme algoritmaları oluşturulduğu için daha az hatalı ve daha güvenilir sonuçlar elde edildiği düşünülmektedir (Strobl, Malley ve Tutz, 2009). Daha sonra sırasıyla Ridge Lojistik Regresyon, Lojistik model, Hoeffding Tree, J4.8, RepTree, Decision Stump ve son olarak Random Tree yöntemlerinin en az hataya sahip ve en güvenilir yöntemler olduğu belirlenmiştir. Bunun yanında tahminleme yöntemlerinin geçerlik ölçütlerinden doğru pozitif, yanlış pozitif, duyarlık, geri çağırma, f-değeri, MKK, ROC eğrisi altında kalan alan ve Duyarlık-geri getirme (DG) eğrisine göre en iyi yöntemin Random forest olduğu belirlenmiştir. Daha sonra geçerlik ölçütlerine göre sırasıyla Ridge Lojistik Regresyon, Lojistik model, Hoeffding Tree, J4.8, RepTree, Decision Stump ve son olarak Random Tree yöntemlerinin daha başarılı oldukları belirlenmiştir. Elde edilen sonuçlar birlikte değerlendirildiğinde yöntemlerin güvenilirlik ve geçerlik ölçütleri bakımından başarı sırasında bir farklılık olmadığı, her bir yöntemin kendi içinde kararlı olduğu belirlenmiştir. Alanyazında en çok kullanılan yöntem olarak görülen J4.8 algoritması orta düzeyde bir başarı göstermiştir. Bunun yerine Random forest, Ridge lojistik regresyon, lojistik model ve Hoeffding tree yöntemlerinin hem hataya dayalı güvenilirlik değerleri hem de geçerlik ölçütleri bakımından daha başarılı yöntemler olduğu belirlenmiştir.

İkinci Alt Probleme İlişkin Sonuçlar

Çalışmanın ikinci alt probleminde farklı yöntemler yardımıyla test edilen verilerin %5, %10, %15, %20, %25, %30 ve %35 olması durumunda elde edilen sonuçların güvenilirlik ve geçerlik değerleri karşılaştırılmıştır. Sinha ve May (2005) bekletme yaklaşımı olarak da bilinen eğitim ve test veri seti yönteminin model performansını tahmin etmede büyük örneklem gruplarında başarılı olurken görece olarak daha küçük gruplarda bu yöntemin normalden daha iyimser sonuçlar vereceğini belirtmektedir. Bu sınırlılığı ortadan kaldırmak için tesadüfi olarak belirlenen eğitim ve test veri setlerinin ortalaması alınarak elde edilen çapraz geçişleme yönteminin kullanılması önerilmektedir (Weiss ve Kulikowski, 1991). Bu

nedenle bu araştırmada farklı veri setlerinde öğrenme yöntemlerinin nasıl bir performans gösterdiği incelenmiş ve doğru pozitif oranları bakımından en iyi sonuçların sırasıyla Ridge lojistik regresyon, Random forest, Lojistik model, Decision stump, Hoeffding tree, J4.8, Rep tree ve Random tree yöntemleri tarafından elde edildiği belirlenmiştir. Ortalama hataların karekökleri esas alınarak karşılaştırma yapıldığında en düşük hata değerlerinin tüm örneklem gruplarında sırasıyla Ridge lojistik regresyon, Random forest, Lojistik model, Hoeffding tree, Decision stump, J4.8, Rep tree ve Random tree yöntemleri tarafından elde edildiği belirlenmiştir. Yöntemlerin farklı örneklem gruplarında nasıl bir değişim gösterdiği incelendiğinde Decision stump yönteminin test edilen veri seti %15 ve üzerinde olması durumunda daha kararlı bir kestirim yaptığı saptanmıştır. Hoeffding tree yönteminin test edilen veri seti %20'ye kadar olması durumunda kararlı sonuçlar üretirken bu değerden sonra hata değerlerinin artarken kappa istatistiğinin azaldığı belirlenmiştir. J4.8 algoritmasının hata değerlerinin farklı örneklemelerde hem arttığı hem de azaldığı görülmüş ve elde edilen sonuçların kararlı olmadığı saptanmıştır. Lojistik modelde test edilen veri setinin hata ve kappa istatistiğinde anlamlı düzeyde bir farklılığa neden olmadığı ve yöntemin oldukça kararlı kestirim yaptığı belirlenmiştir. Rep tree yönteminin de hata değerlerinin farklı veri setlerinde hem arttığı hem de azaldığı görülmüş ve elde edilen sonuçların kararlı olmadığı saptanmıştır. Random forest yöntemi farklı veri setlerinde hata değerleri bakımından kararlı kestirimlerde bulunurken sadece kappa istatistiğinin %15-%25 aralığında bir miktar arttığı sonra tekrar aynı düzeye indiği belirlenmiştir. En yüksek hata değerlerine sahip olan Random tree yönteminin test edilen veri setinin %15 oluncaya kadar nispeten kararlı kestirimlerde bulunurken sonrasında test edilen ölçütlerde artma ve azalmalar olduğu görülmüştür. En iyi kestirimde bulunan Ridge regresyon yönteminin tüm test verilerinden oldukça kararlı kestirimlerde bulunduğu belirlenmiştir.

Üçüncü Alt Probleme İlişkin Sonuçlar

Çalışmanın üçüncü alt probleminde farklı yöntemler yardımıyla elde edilen ölçme sonuçlarının veri setini test ve eğitim verisi olarak ayırmayıp aynı veri seti üzerinden hem öğrenme yöntemini eğitip hem de test ettiğimizde karışıklık matrisi yardımıyla elde edilen sonuçların doğruluk dereceleri karşılaştırılmıştır. Buna göre Random forest ve Random tree yöntemlerinin tüm örnekleri doğru olarak sınıflayarak en yüksek sınıflama oranına sahip olduğu belirlenmiştir. Üçüncü en yüksek doğru sınıflama oranına sahip J4.8 algoritmasını sırasıyla Rep Tree, Lojistik model, Ridge regresyon, Hoeffding tree ve Decision stump yöntemleri takip etmektedir. Aynı veri seti üzerinde kestirimde bulunduğu taktirde aslında hatalı ve düşük doğru sınıflama oranlarına sahip olan Random tree ve Rep tree yöntemlerinin %100 gibi oldukça başarılı birer algoritma olarak belirlenmesi oldukça düşündürücüdür. Bu sorun veri madenciliğinde aşırı uyum gösterme (overfitting) olarak tanımlanan sorundur (Domingos, 2012). Her ne kadar aşırı uyum gösterme sorunu yanlış kestirim ve yüksek varyans gibi farklı şekillerde ortaya çıkabilse de aynı veri seti üzerinde çalışmak bu soruna neden olabilmektedir (Domingos, 2000). Bu sorunu ortadan kaldırmanın en iyi yöntemlerinde biri de çapraz geçişleme uygulamaktır (Ng, 1997). Araştırmada çapraz

geçerleme yöntemi uygulandığında aşırı uyum gösteren Random Tree ve RepTree yöntemlerinin genel başarı oranlarının oldukça düştüğü belirlenmiştir. Bu sonuca göre alanda çalışma yapacak araştırmacıların topladıkları verileri eğitim ve test veri seti olarak ayırmamaları halinde kullanacakları yöntemlerin hatalı sonuçlar vereceği görülmektedir. Bu sebeple araştırmacıların veri madenciliği alanında çalışma yaparken test edilen veri setini en az %15 ve üzerinde tutmaları gerekmektedir. Bunun yanında daha kararlı ve daha doğru kestirimlerde bulunmak için çapraz geçerleme yöntemlerini kullanmaları ve bu aşamada kullanacakları katman sayısının en az 10 ve üzeri olması önerilmektedir.

Model değerlendirme ölçütlerinden ROC eğrisi altında kalan alanların incelenmesinde Elayidom (2012) tarafından yapılan değerlendirme kriterleri esas alınmıştır. Buna göre eğri altındaki alan 1.00-0.90 arası mükemmel, 0.80-0.89 arası iyi, 0.70-0.79 arası normal, 0.60-0.69 arası zayıf ve 0.50-0.59 arası başarısız olarak yorumlanmaktadır. Çalışmada farklı algoritmalar yardımıyla 10 katlı çapraz geçerleme sonucunda sırasıyla Random forest, lojistik model, Ridge lojistik regresyon, Hoefding tree ve Rep tree yöntemlerinin başarılı ve başarısız olarak tanımlanan öğrencileri normal seviyede tahmin ederken J4.8, Random tree ve Decision stump yöntemlerinin zayıf olarak tanımlanan seviyede kestirimde buldukları belirlenmiştir. Mehdiyev, Enke, Fettke ve Loos (2016) tarafından yapılan benzer bir çalışmada en doğru kestirimlerin sırasıyla yapay sinir ağları, Random forest, Lojistik regresyon, Radyal temelli ağlar ve C4.5 yöntemleri tarafından gerçekleştirildiği görülmektedir. Elde edilen bu sonuç araştırmanın bulgularıyla benzerlik göstermektedir. Sonuç olarak Weka programında yer alan sekiz farklı öğrenme yönteminin farklı değerlendirme ölçütleri altında en iyi sonuçlar Random forest, Ridge regresyon yöntemi ve lojistik model tarafından elde edilmiştir.

Öneriler

Araştırmadan elde edilen bulgulara dayalı olarak aşağıdaki önerilerde bulunulmuştur.

1. Değişken sayısını azaltmak ve en çok bilgi veren özellikleri belirleme amacıyla farklı yöntemlerle 10 katlı çapraz geçerleme yaklaşımı ile çalışma kapsamında kullanılacak değişken sayısı 29'dan 12'ye indirilmiştir. Elde edilen bulgulara dayalı olarak fen başarısının yordanması amacıyla sınırlı sayıda bağımsız değişken almak yerine çok sayıda değişkenin farklı katmanlardaki doğru tahminleme başarıları incelenerek daha tutarlı ve güvenilir kestirimler elde edilebilir.
2. Çalışma kapsamında elde edilen sonuçların güvenilirliğine ilişkin delil sunmak amacıyla mutlak ve bağıl hata değerlerinin farklı birimlerle ifade edildiği belirlenmiştir. Elde edilen bu bulguya dayalı olarak veri madenciliğinde elde edilen sonuçların güvenilirliğine ilişkin bilgi vermek amacıyla ortalama hataların karekökünün yanında standartlaştırılmış hata değerlerinin rapor edilmesi önerilmektedir.

3. Veri madenciliğinde örneklemin yeterince büyük olması ve üzerinde çalışılan öğrenme yöntemini aynı veri seti üzerinde eğitmek ve test etmek yerine çapraz geçişleme gibi yöntemlerin kullanılmasının gerektiği görülmektedir. Aksi takdirde aşırı uyum gösterme sorunuyla karşılaşılacağı belirlenmiştir. Bu sebeple aynı veri setinin hem test hem de eğitim için kullanılmaması önerilmektedir.
4. Çalışmadan elde edilen güvenilirlik ve geçerlik ölçütlerinin farklı örneklem büyüklüklerinde değişkenlik gösterdiği belirlendiğinden örneklemin 1000 ve üzerinde olması durumunda test edilen veri setinin en az %15 ve üzeri olması gerekmektedir. Aksi takdirde elde edilen sonuçların tutarlı olmayacağı göz önünde bulundurulması gerekir.
5. Veri madenciliğinde elde edilen sonuçları sadece karışıklık matrisinden elde edilen doğruluk değerlerine göre yorumlamak yerine duyarlılık, seçicilik, ROC eğrisi altında kalan alan, Matthew korelasyon katsayısı, F-ölçütü ve Geri çağırma gibi standart ölçütlerin de mutlaka raporlanarak elde edilen sonuçların güvenilirlik ve geçerliğine ilişkin birden fazla değer rapor edilmesi önerilebilir. Bu sayede güvenirlige ve geçerliğe ilişkin birden fazla delil sunulmuş olacaktır.
6. Veri madenciliğinde özellikle Weka programında yer alan öğrenme yöntemlerinden Random forest, lojistik model (LMT) ve Ridge regresyon yöntemlerinin farklı koşullarda altında genellikle en güvenilir ve geçerli sonuçları elde ettiği belirlendiğinden başarının yordanması veya sınıflama amacıyla yapılacak çalışmalarda bu öğrenme yöntemlerinin kullanılması önerilmektedir.
7. Özellikle lojistik modelde başarılı ve başarısız olarak sınıflanan öğrencilere ilişkin elde edilecek modelde değişkenler arasındaki ilişkinin tek bir regresyon denklemi ile modellenmesi yerine farklı özellik değerlerine göre çok sayıda regresyon denklemi kullanarak doğru sınıflama oranı artırılabilir.
8. Bekletme yönteminde test veri setinin en az %15 ve üzeri olması durumunda geçerlik ve güvenilirlik ölçütlerinin en yüksek değere ulaşması sebebiyle bu alanda çalışma yapacak araştırmacıların daha düşük orandaki test veri setlerinden elde edecekleri sonuçlara dikkat etmeleri önerilmektedir.

Sınırlılıklar

Araştırmanın sınırlılıkları aşağıda verilmiştir.

1. Veri madenciliğinde kullanılan güvenilirlik ölçütlerinin göreceli ve yüzdelerle ifade edilmesi sebebiyle bu ölçütlerin kıyaslanabilir nitelikte olmadığı belirlenmiştir. Bu nedenle standart ölçütlerin araştırmacılar tarafından daha kolay anlaşılabilmesi düşünüldüğünden daha farklı güvenilirlik ölçütlerinin belirlenmesinin gerektiği düşünülmektedir.

2. Çalışmada model değerlendirme yaklaşımlarından 10 katlı çapraz geçirme yönteminin kullanılması araştırmanın sınırlılığı olarak görülmektedir. Bunun yerine $k=5, 10, 15$ ve 20 alınarak elde edilen sonuçlar karşılaştırılarak daha kararlı bulgular elde edilebilir.

Kaynakça

- Ahmed, A. B. and Elaraby, I. S. (2014) Data mining: A prediction for student's performance using classification method. *World Journal of Computer Application and Technology*, 2(2), 43-47.
- Boss, D. D. (2003). Introduction to the bootstrap world. *Statistical Science*, 18(2), 168-174.
- Bramer, M. (2013). *Principles of data mining (2nd ed.)*. London: Springer-Verlag.
- Brown, M. S. (2014). *Data mining for dummies*. Hoboken, New Jersey: John Wiley and Sons.
- Chamatkar, A. J. and Butey, P. K. (2014). Importance of data mining with different types of data applications and challenging areas. *Journal of Engineering Research and Applications*, 4(5), 38-41.
- Chen, S. X. and J. S. Liu (1997). Statistical applications of the Poisson-binomial and conditional Bernoulli distributions. *Statistica Sinica* 7, 875–892.
- Dekking, F. M., Kraaikamp, C., Lopuhaa, H. P. and Meester, L. E. (2005). *A modern introduction to probability and statistics: understanding why and how*. United States of America: Springer Science+Business Media.
- Domingos, P. (2012). A few useful things to know about machine learning. *Communications of the ACM*, 55(10), 78–87.
- Egan, J.P. (1975). *Signal detection theory and ROC analysis*. Series in Cognition and Perception. Academic Press, New York.
- Efron, B. (1979). Bootstrap methods: another look at the jackknife. *Annals of Statistics*, 7, 1–26.
- Elayidom, S. M. (2012). *Design and development of data mining models for the prediction of manpower* (Yayınlanmamış doktora tezi). Cochin University of Science and Technology Computer Science and Engineering, Kochi, India.
- Elhamahmy, M. E., Elmahdy, H. N. and Saroit, I. A. (2010). A new approach for evaluating intrusion detection system. *CiiT International Journal of Artificial Intelligent Systems and Machine Learning*, 2(11), 290-298.
- Fernandez-Delgado, M., Cernadas, E., Barro, S. and Amorim, D. (2014). Do we need hundreds of classifiers to solve real world classification problems? *Journal of Machine Learning Research*, 15, 3133–3181.

- Fraenkel, J. R., Wallen, N. E., and Hyun, H. H. (2012). *How to design and evaluate research in education (8th ed.)*. New York: Mc Graw Hill.
- Friedman, J. H., and Fisher N. I. (1999). Bump hunting in high-dimensional data. *Stat Comput*, 9,123–143.
- Galdi, P., and Tagliaferri, R. (2017). *Data mining: Accuracy and error measures for classification and prediction*, in Reference Module in Life Sciences, Holland: Elsevier.
- Han, J., and Kamber, M. (2006). *Data mining: Concepts and techniques*, (2nd edition).
- Hastie, T., Tibshirani, R., and Friedman, J. H. (2009). *The elements of statistical learning: Data mining, inference, and prediction*. New York, NY: Springer.
- Huang, S., and Fang, N. (2013). Predicting student academic performance in an engineering dynamics course: A comparison of four types of predictive mathematical models. *Computers and Education*, 61, 133–145.
- Kohavi, R. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. *Appears in the International Joint Conference on Artificial Intelligence*, 2, 1137-1143.
- Kumar, R., and Indrayan A. (2011). Receiver operating characteristics (ROC) curve for medical researchers. *Indian Pediatrics*, 48, 277-287
- Kuonen, D. (2018). An introduction to bootstrap methods and their application. *WBL in Angewandter Statistik ETHZ 2017/19*, 1-143.
- Liaw A., and Wiener M. (2002). Classification and regression by random forest. *R News*, 2 (3), 18-22.
- Lykourentzou, I., Giannoukos, I., Mpardis, G., Nikolopoulos, V., and Loumos, V. (2009). Early and dynamic student achievement prediction in e-learning courses using neural networks. *Journal of the American Society for Information Science and Technology*, 60(2), 372–380.
- Mavroforakis, C. (2011). *Data mining with WEKA*. Boston University. http://cs-people.bu.edu/cmav/cs105/files/lab12/intro_to_weka.pdf adresinden elde edilmiştir.
- MEB (2015). PISA 2015 Ulusal Raporu. Ankara: MEB Ölçme Değerlendirme ve Sınav Hizmetleri Genel Müdürlüğü.
- MEB (2016). PISA 2015 Ulusal Ön Raporu. Ankara: MEB.
- Mehdiyev, N., Enke, D., Fettke, P., and Loos, P. (2016). Evaluating forecasting methods by considering different accuracy measures. *Procedia Computer Science*, 95, 264 – 271.

- Ng, A.Y. (1997). Preventing "overfitting" of cross-validation data. In D.H. Fisher (Ed.), *Proceedings of the 14th International Conference on Machine Learning*, Nashville, TN, USA, July 8–12, San Francisco, CA: Morgan Kaufmann.
- North, M. A. (2012). *Data Mining for the Masses*, ABD: A Global Text Project Book.
- Olmo, J.L. Romero, J.R., and Ventura, S. (2012). Classification rule mining using ant programming guided by grammar with multiple Pareto fronts. *Soft Computing*, 16(12), 2143-2163.
- Ramageri, M. B. (2010). Data mining techniques and applications, *Indian Journal of Computer Science and Engineering*, 1(4), 301-305.
- Refaeilzadeh, P., Tang, L., and Liu., H. (2009). *Cross validation*. In *Encyclopedia of database systems*, (Editors: M. Tamer A-zsu and Ling Liu). New York, ABD: Springer publishing.
- Schwenke, C., and Schering, A. (2007). *True positives, true negatives, false positives, false negatives*. New Jersey, ABD: Wiley Encyclopedia of Clinical Trials.
- Sinha, A. P., and May, J. H. (2005). Evaluating and tuning predictive data mining models using receiver operating characteristic curves. *Journal of Management Information Systems*, 21(3), 249-280.
- Sinharay, S. (2016). An NCME instructional module on data mining methods for classification and regression. *Educational Measurement: Issues and Practice*, 35 (3), 38–54.
- Souza, J., Matwin, S., and Japkowicz, N. (2002). *Evaluating data mining models: a pattern language*. In: 9th Conference on Pattern Language of Programs (PLOP'02), Monticello, Illinois, 8–12 September 2002.
- Strobl, C., Malley, J., and Tutz, G. (2009). An introduction to recursive partitioning: Rationale, application, and characteristics of classification and regression trees, bagging, and random forests. *Psychological Methods*, 14, 323–348.
- Svetnik, V., Liaw, A., Tong, C., and Wang, T. (2004). Application of Breimans random forest to modeling structure-activity relationships of pharmaceutical molecules. In F. Roli, J. Kittler ve T. Windeatt (Eds.), *Multiple classifier systems* (cilt 3077, 334–343). Cagliari, Italy: Springer.
- Tatlı, C. E., Ergin, D. A. ve Demir, E. (2016). PISA 2012 Türkiye verilerine göre öğrencilerin matematik kaygısının sınıflandırıcıları. *İlköğretim Online*, 15(2), 696-707.
- Vanwinckelen, G., and Blockeel, H. (2012) *On estimating model accuracy with repeated cross-validation*, Belgian-Dutch Conference on Machine Learning (BeneLearn) edition:21 location, 24-25 May 2012.

- Weiss, S. M., and Kulikowski, C. A. (1991). *Computer systems that learn: Classification and prediction methods from statistics, neural nets, machine learning, and expert systems*. San Mateo, CA: Morgan Kaufmann.
- Williams, G. (2011). *Data mining with Rattle and R: The art of excavating data for knowledge discovery*, New York, USA: Springer Science+Business Media.
- Witten, I. H., and Frank, E. (2005) *Data mining: Practical machine learning tools and techniques*, United States of America: Morgan Kaufmann publications
- Witten, I. H., Frank, E., and Hall, M. (2016) *Data mining: Practical machine learning tools and techniques*, United States of America: Morgan Kaufmann publications.



Comparison of Learning Methods Used in Data Mining Under Different Conditions ¹

ARTICLE TYPE	Received Date	Accepted Date	Online First Date
Research Article	09.26.2018	11.10.2018	11.12.2018

Gökhan Aksu ²

Adnan Menderes University

Nuri Doğan ³

Hacettepe University

Abstract

In this study, it is tried to determine the level of reliability and validity of data mining and machine learning approach in the field of education and the results obtained based on these algorithms. In the study according to the 2015 PISA average of Turkey in terms of scientific literacy, students were classified as failing and successful using different learning methods and the reliability and validity measures of the results obtained at this stage were examined. The best results in terms of the correct classification rate, kappa statistic, square root error and relative square root error values were obtained from Random Forest method besides this ridge logistic regression, logistic model and Hoefding tree methods were determined as the most successful other methods. In the case that the whole data set is separated as a training and test data set without using the cross validation method, it was found that the Random Tree and J.48 methods had the highest error values while the logistic model, Random Forest and Ridge Regression methods gave the lowest error values in different sized test data. It is concluded that the error values obtained by Ridge regression, Random forest and Logistic model are quite consistent in different percentage test data. It was determined that the measurement results obtained with the help of different methods did not divide the data set as test and training data, but if we train both the learning method and the same data set through the same data set, Random tree and J.48 learning methods had a higher accurate classification rate than the real performances.

Keywords: Data mining, WEKA, learning method, classification, PISA.

¹This study was produced form a PhD thesis entitled "Comparison of Learning Methods Used in Data Mining Under Different Conditions" carried out by Aksu (2018).

²*Corresponding Author:* PhD, Adnan Menderes University, Vocational High School, E-mail: gokhanaksu1983@hotmail.com, <https://orcid.org/0000-0003-2563-6112>

³Prof. Dr., Faculty of Education, Deparmant of Education, E-mail: nuridogan2004@gmail.com, <https://orcid.org/0000-0001-6274-2016>

Purpose and Significance

The day of the information age, data is collected from a wide variety of sources. With the help of this information, it is estimated about their behavior based on many different characteristics of individuals. At this stage, data mining methods are used especially in recent years. In this study, it is aimed to determine how the reliability and validity values of the results obtained by different learning methods used in data mining vary under different conditions. Within the scope of this aim, "What are the reliability and validity values of the data mining classification algorithms used to predict the success of the students who participated in the PISA exam in terms of science literacy?" was sought. In the first sub-problem of the study, the reliability and validity values which estimating the science achievement obtained with the Regression methods of Decision Stump, Hoeffding Tree, J.48, Logistic Model, RepTree, Random Forest, and Logistic Model, with the help of 10-fold cross validation method were determined. In the second sub-problem the results obtained with the help of Decision Stump, Hoeffding Tree, J.48, Logistic Model, RepTree, Random Forest, Random Tree and Ridge Logistic methods tested in 5%, 10%, 15%, 20%, 25%, 30% and 35% because of determining the level of reliability and validity. In the third sub-problem the degree of accuracy obtained with the help of the confusion matrix was determined with the help of Decision Stump, Hoeffding Tree, J.48, Logistic Model, RepTree, Random Forest, Random Tree and Ridge Logistic methods.

Method

The WEKA program was used in the study in which the data mining methods were used to determine the reliability and validity values of the data mining estimation methods used in this process. For the purposes of the study, the study group consisted of 15 age group students who participated in the PISA 2015 exam organized by OECD and registered in formal education. Examination total of around 540,000 students from 72 countries participated and 5895 were from Turkey. The universe of Turkey PISA 2015 exam is 1.324.089 and the available sample of Turkey is 925.366 students. The data used in the study was obtained from the official web site of OECD, which was published in 2017 and was published on the <http://www.oecd.org/pisa/data/2015database/> OECD website. In the research, use of information technology, student's past science experience, the support provided for family learning, the number of schools where students change and family perceive that school quality variables are could not be included in the analysis due to the questions left blank and / or not answered. In the 10-fold cross validation analysis, it was determined that the variables had a significant effect on at least 4 layers. Accordingly, the number of input variables total of 12, covered in the study are based on inquiry Science education, sensitivity to the environment, scientific beliefs, expected professional status from students, out-of-school study time (weeks), duration of science learning (week), total learning period (week), test anxiety, fairness of the teacher, educational resources at home, ICT resources and socioeconomic status index. PISA science literacy scores of students in the first phase of the research, which

is below the average in Turkey fails 425.00 points (0), is accepted as successful and this score as coded 1. According to the results of the two categories, 2769 of the 5865 students were classified as successful (47.20%) and 3096 (52.8%) of the students as failed. In the second phase of the study, a model was created to predict students' PISA science literacy performance by using independent variables (Hastie, Tibshirani and Friedman, 2009). Because students' performance in the PISA exam is coded as Successful and Unsuccessful, this is a classification problem and estimations have been made using data mining classification methods (Olmo, Romero and Ventura, 2012). For this purpose, Decision Stump, Hoeffding Tree, J.48, Logistic Model, RepTree, Random Forest, Random Tree and Ridge Logistic Regression methods were used. In the third phase of the study; The reliability and validity values of the students classified as successful and unsuccessful with the classification algorithms determined within the scope of the research were examined. Reliability criteria as correct classification rate, correct classification number, Kappa statistic, mean absolute error, square root error, relative absolute error and relative square root error values, and validity criteria as correct criterion, false positive rate, sensitivity, recall, F- criterion, Matthew correlation coefficient, area under ROC curve and Area-Sensitivity-Recall curve are used. In order to determine the external validity of the results obtained from the Weka program, analyzes were performed again in the Matlab program and the correct classification rates of both programs were compared over the same training and test data set. 66.6% of the data obtained from 5865 students who participated in the PISA 2015 exam was divided into training (N = 3870) and 33.3% (n = 1995) as the test data set. The overall success rate was determined as 71,03% (2749/3870) by 1392 students who failed and 1357 students who were successful in the realization by classifying method in the Weka program. Following the introduction of the same data into the Matlab2017b program as a set of training and test data, the overall success rate was determined to be 70.31% (2721/3870), with 1039 successful individuals who failed in the training data set and were actually successful by the classification method. Overall the correct classification rate of Weka program was 71,03% and 69,90% in Matlab program.

Results

Weighted general average values were compared in order to compare the reliability values obtained with different methods in the first sub-problem of the study. Accurate classification rates for each method, correct classification rates, Kappa statistics, average of absolute error, mean square root of errors and relative absolute error values were determined. Accordingly, the best results in terms of the correct classification, correct classification rate, kappa statistics, square root error and relative square root error values were obtained by the Random Forest method. Besides this, it was determined that the best results were obtained by Hoeffding Tree method in terms of the mean of absolute error and relative absolute error. In addition to this, while the number and number of samples correctly classified, the Kappa statistics, the mean square root of the errors and the square root of the relative errors, the worst results are obtained by the Random tree. It was determined that the worst results were obtained

by Decision Stump method in terms of the mean of absolute error and relative absolute error. When the validity criteria obtained by using different methods were compared, it was determined that all validity criteria obtained by Random Forest method were higher than other methods. When the results obtained for reliability and validity are evaluated as a whole it is seen that the best results are random forest, Ridge logistic regression, Logistic model, Hoefding tree, J.48, RepTree, Decision Stump and Random Tree algorithms, respectively. If you are asked to make a ranking considering the Kappa statistic it is seen that the best results will be in the form of Random forest, Ridge logistic regression, Logistic model, Hoefding tree, J.48, RepTree, Decision Stump and Random Tree algorithms respectively. Accordingly, it was determined that using the correct classification or kappa statistic would not cause a change in the ranking.

In the second sub-problem of the study, it was determined that the reliability and validity results of different methods were 5%, 10%, 15%, 20%, 25%, 30% and 35%. According to this, the logistic model, Random Forest and Ridge Regression methods gave the most accurate estimation results in different sized test data and it was determined that the data set rate was 15% when it reached the highest value and then decreased again. Besides, Logistic model, Random Forest and Ridge Regression methods have the lowest error values in different size test data. It is observed that the lowest error values are obtained when the rate of the data set is 15% and then increased again.

In the third sub-problem of the study, it was determined that the level of accuracy calculated by the confusion matrix of different methods. According to this, Random Forest and Random Tree methods were determined to have a 100% accurate classification rate. Following these methods, it was determined that the learning method with the best correct classification ratio was J.48 algorithm and the success rate was 80.95%. Decision Stump method's success rate were found to be 63.87%. However via 10-fold cross validation method Random tree with a success rate of 67.28% and J.48 methods with a success rate of 67.79%, if the data set is not separated as education and test data, it is seen that they have a higher correct classification rate as 100% and 80.95% respectively.

Discussion and Conclusions

It is determined that the best estimation method according to the square root of the categorized samples, kappa statistic, absolute error and mean errors is Random forest. In the random forest method, it is thought that the use of multiple classifiers instead of a classifier and then the results obtained from their estimates and learning algorithms that classify the new data are generated, and therefore less faulty and more reliable results are obtained (Strobl, Malley and Tutz, 2009). The J4.8 algorithm, which is the most commonly used method in the literature, showed a moderate success. Instead, Random forest, Ridge logistic regression, logistic model and Hoefding tree methods were found to be more successful methods in terms of error-based reliability values and validity criteria.