Öğrenci Proje Anketlerini Sınıflandırmada En Başarılı Algoritmanın Belirlenmesi

Determining The Most Successful Classification Algorithm For The Student Project Questionnaire

Pınar Cihan[1], Oya Kalıpsız[2]
[1]Yıldız Teknik Üniversitesi, pinar@ce.yildiz.edu.tr
[2]Yıldız Teknik Üniversitesi, kalipsiz@yildiz.edu.tr

Öz

Pek çok alanda etkili bir şekilde kullanılan veri madenciliğinin, günümüzde eğitim alanındaki uygulamaları hızla artmaktadır. Veri madenciliği yöntemleri ile eldeki veriler sınıflandırılarak, gruplandırılarak ya da veriler arasında ilişkiler, bağıntılar, istatistiksel sonuçlar oluşturularak modeller oluşturulur. Oluşturulan model, oluşturulduğu veri kümesinde olmayan yeni bir kayıt geldiğinde, yeni gelen kayıt hakkında tahminleme yapar. Yapılan tahminlerin doğruluk derecesi oluşturulmuş modelin veri üzerindeki başarısını ortaya koyar. Bu çalışmada Yıldız Teknik Üniversitesi(YTU), Bilgisayar Mühendisliği bölümü, Sistem Analizi ve Tasarımı dersinde gerçekleştirilen projeler ile ilgili öğrencilere yapılan anketlerden veri seti oluşturulmuştur. Elde edilen bu veri setine sınıflandırma algoritmaları uygulanarak en başarılı algoritma tespit edilmiştir. Çalışmadaki amaç proje tabanlı gerçekleştirilen derslerde, proje başarısını arttırmak için öğrenci projelerinin sınıflandırılmasında en başarılı sınıflandırma algoritmasının tespit edilmesidir. En başarılı algoritma tespit edilirken Doğruluk, Ortalama Mutlak Hata(MAE), Kök Hata Kareler Ortalaması(RMSE) ve Kappa değerleri göz önüne alınmıştır. Sonuç olarak öğrenci projelerinden elde edilen veri setini sınıflandırmada en başarılı algoritmanın Çok Katmanlı Algılayıcı(MLP) olduğu tespit edilmiştir. Ayrıca Apriori Algoritması kullanılarak bazı kurallar elde edilmiştir.

Anahtar Sözcükler: *Veri Madenciliği, Eğitimde Veri Madenciliği, Veri Sınıflandırma, Sınıflandırma Algoritmaları, Birliktelik Kuralı, Apriori Algoritması.*

Abstract

Data mining is used in many areas including knowledge discovery tasks in machine learning, statistics, and database systems. At present particularly in educational area, use of data mining and its applications is increasing. Data mining methods are used to create models from the data by classification, clustering or simply applying correlation rules in the data. When the new instance/record, which is not available in the current data set, appears, this model can be used to make assumptions regarding to that instance. Precision level of those estimates shows the accuracy of the created model on the data set. In this study, the data set is created by using questionnaires, which were collected from System Analysis and Design courses at Computer Engineering Department, Yıldız Technical University(YTU). Those questionnaire consist of student answers that are related to the projects, which are developed in these courses. Classification algorithms are applied on this data set to decide the most successful algorithm for classification. The main goal of this study is to reveal the most successful classification algorithm in terms of student projects classification tasks for purpose of increasing the success rate of the projects in the courses that are based on the projects. Accuracy, Mean Absolute Error(MAE), Root Mean Square Error(RMSE), and KAPPA metrics are used to evaluate performance of those algorithms. The experimental result of this study shows that Multilayer Perceptron(MLP) algorithm performs better that other algorithms on this data set. Also Appriori algorithm used for finding some rules.

Genderim ve kabul tarihi : 25.08.2015-29.09.2015

Keywords: *Data Mining, Educational Data Mining, Data Classification, Classification Algorithms, Association Rule, Apriori Algortihm.*

1. Giriş

Yazılımın hayatımızdaki artan önemi yazılım mühendisliğine olan ilgiyi hem akademik çevrelerde hem de endüstride artırmıştır. Günümüzde yazılım mühendisliği yazılım sektöründe en zorlu işlerden biridir ve yazılımlarda karşılaşılan problemler, yazılım mühendisliği eğitiminin önemini vurgulamaktadır. Bu nedenle, yazılım problemlerinin çözümü ve iyileştirmeler için, yazılım mühendislerinin eğitimine odaklanılması gerektiği vurgulanmaktadır [1,2].

Endüstrideki düşük başarı oranlı yazılım projeleri çok ciddi maliyete sebep olmaktadır. Bu nedenle yazılım mühendisliği eğitimi, öğrenciler mezun olmadan ve önemli tasarım ve uygulama sorumlulukları almadan önce mutlaka bazı pratik deneyimleri kazandırmalıdır [1].

Aynı zamanda Veri Madenciliği, günümüzün en çok uygulanan disiplinlerinden birisidir. Her geçen sene kendisine daha da yaygın bir kullanım alanı bulmakla birlikte, kolay uygulanabilirliği ve etkili sonuçlar ortaya çıkarması sayesinde en çok başvurulan yöntemlerden bir tanesidir. Literatürde veri madenciliği eğitim, ticaret, mühendislik, bankacılık ve borsa, tıp ve telekomünikasyon gibi birçok alanda kullanılmaktadır. Eğitim alanında çok sayıda gerçekleştirilen veri madenciliği uygulamalarının bir kaçı şunlardır:

Kurt ve Erdem çalışmasında [3], öğrencilerin başarılarına etki edebilecek faktörleri farklı veri madenciliği algoritma ve modelleriyle incelenmiştir. Ekonomik, sosyal, kişisel, çevresel değişkenler üzerindeki yapılan uygulamada çeşitli sonuçlar saptanmış, uygun öneriler sunulmuştur. Bırtıl tezinde [4], kız meslek lisesi öğrencilerinin akademik başarısızlık nedenlerini belirlemek için öğrencilere uygulanan anketler veri madenciliği yöntemi kullanılarak değerlendirilmiş ve öğrencileri başarısızlığa iten etkenlerin hangilerinin aynı anda görüldüğü ve aralarındaki ilişkiler tespit edilmiştir. Bozkır vd. tarafından gerçekleştirilen çalışmada [5], ÖSYM tarafından 2008 ÖSS adayları için resmi internet sitesi üzerinden yapılan anket verileri üzerinde veri madenciliği yöntemleri kullanılarak, öğrencilerin başarılarını etkileyen faktörler araştırılmıştır. Bu araştırmada, veri madenciliği yöntemlerinden karar ağaçları ve kümeleme kullanılmıştır. Aydın tezinde [6], Anadolu Üniversitesi Uzaktan Eğitim Sisteminde eğitim gören öğrencilere ilişkin farklı kaynaklardaki verileri bir araya getirerek veri madenciliği uygulaması gerçekleştirmiştir. Öğrenci başarısını tahmin etmeye yönelik çalışmada C5.0 karar ağacı algoritmasının kullanıldığı bir tahmin modeli önerilmiştir. Ayık vd. tarafından yapılan çalışmada [7], Atatürk Üniversitesi öğrencilerinin mezun oldukları lise türleri ve lise mezuniyet dereceleri ile kazandıkları fakülteler arasındaki ilişki, veri madenciliği teknikleri kullanılarak incelenmiştir. Karabatak ve İnce çalışmasında [8], Veri Madenciliği teknikleri kullanılarak Fırat Üniversitesi Teknik Eğitim Fakültesi Bilgisayar Eğitimi bölümü öğrencilerinin notları kullanılarak öğrenci başarılarının analizi yapılmıştır. Veri madenciliği tekniği ile öğrencilerin dersleri ile notları arasındaki ilişkiler ortaya çıkarılmıştır.

Veri madenciliği konusunda çok sayıda yöntem ve algoritma geliştirilmiştir. Veri madenciliği modelleri temel olarak sınıflandırma, kümeleme ve birliktelik kuralları şeklinde sınıflandırılabilir. Bu çalışmada YTU, Sistem Analizi ve Tasarımı dersinde proje gerçekleştiren 86 öğrenciye projeleri ile ilgili anket uygulandıktan sonra 20 anket sorusu üzerinde veri madenciliğinde sıkça kullanılan yöntemlerden biri olan sınıflandırma yöntemleri kullanılarak öğrenci projelerini sınıflandırmada(başarılı /başarısız) en başarılı olan sınıflandırma algoritması tespit edilmiştir.

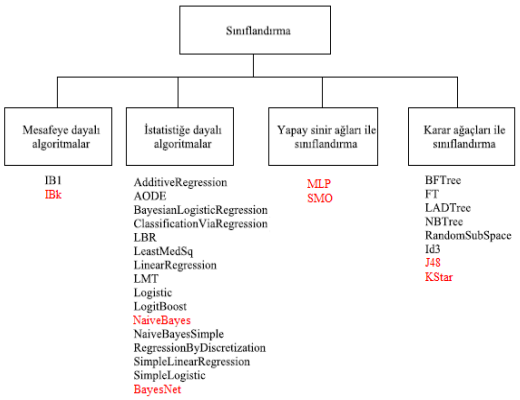
Ayrıca veriler arasındaki beklenmeyen ilişkileri bulmak, gizli bilgileri açığa çıkararak gelecekteki eğilimleri belirlemek için, veri madenciliğinde, birliktelik kuralı çıkarım algoritmalarından biri olan Apriori algoritması kullanılmıştır.

Çalışmada Waikato Üniversitesinde java programlama diliyle geliştirilmiş olan WEKA (Waikato Environment for Knowledge Analysis) programı kullanılmıştır. Sınıflandırmalarda algoritmaların başarıları değerlendirilirken Doğruluk değeri, Ortalama Mutlak Hata(MAE), Kök Hata Kareler Ortalaması(RMSE) ve Kappa istatistiği kıstasları göz önüne alınmıştır.

Elde edilen sonuçlar Bölüm 4'de yer almaktadır.

2. Veri Madenciliği Teknikleri

Sınıflandırma veri madenciliğinde sıkça kullanılan bir yöntemdir. Veri setinde bulunan her örneğin bir dizi niteliği vardır ve bu niteliklerden biri de sınıf bilgisidir. Sınıflandırma bir öğrenme algoritmasına dayanır, hangi sınıfa ait olduğu bilinen veri kümesi (eğitim kümesi) eğitim amacıyla kullanılır ve bir model oluşturulur. Oluşturulan model öğrenme kümesinde yer almayan veri kümesi (deneme kümesi) ile denenerek başarısı ölçülür. Oluşturulan bu model kullanılarak hangi sınıfa ait olduğu bilinmeyen bir kayıt için bir sınıf belirlenebilir. Aşağıda veri madenciliği algoritmaları çerçevesinde WEKA'nın hiyerarşik yapısı ve bu yapı üzerinden kullanılan algoritmalar Şekil 1'de gösterilmiştir.



Şekil-1: Sınıflandırma algoritmasının hiyerarşik yapısı

Çalışma kapsamında mesafeye dayalı algoritmalar ile sınıflandırma, istatistiğe dayalı algoritmalar ile sınıflandırma, yapay sinir ağı ile sınıflandırma ve karar ağaçları ile sınıflandırmada kullanılan bazı algoritmalar incelenmiştir (Tablo 1).

Tablo-1: Kullanılan veri madenciliği teknikleri

Veri Madenciliği Teknikleri	Sınıflandırma Algoritması
Mesafeye Dayalı Algoritmalar	***IBk;*** En yakın K-Komşu (K-Nearest Neighbors) algoritmasıdır. Bu algoritma sınıflandırma için kullanılır. K tabanlı komşuların uygun değerini çapraz doğrulama ile seçebilir. Ayrıca mesafe ağırlıklandırabilir[9].
	KStar; K*, örnek tabanlı bir sınıflandırıcıdır. Bazı benzerlik fonksiyonlarıyla belirlendiği gibi, eğitim örnekleriyle aynı olan sınıfa istinaden, test örneğinin sınıfıdır. Diğer örnek tabanlı öğrenenlerden entropi tabanlı mesafe fonksiyonu kullanması yönüyle farklıdır [10].
İstatistiğe Dayalı Algoritmalar	***Naive Bayes (NB);*** Naive Bayes sınıflandırıcı Bayes teoremine dayanmaktadır. Bu teorem bir rassal değişken için olasılık dağılımı içinde koşullu olasılıklar ile marjinal olasılıklar arasındaki ilişkiyi gösterir. Veri kümesindeki her özelliğin sınıflama problemine eşit katkıda bulunduğu ve katkıların birbirinden bağımsız olduğu varsayıldığında basit bir sınıflama olan NB sınıflayıcısı kullanılabilir. Sınıflandırma yapılırken en yüksek olasılıklı durum hedef sınıf olarak seçilir. ***Bayes Net;*** Bayes ağların özelliği istatistiksel ağlar olmaları ve düğümler arası geçiş yapan kolların istatistiksel kararlara göre seçilmesidir. Bayes ağları yönlü dönüşsüz ağlardır (directed acyclic network) ve her düğüm ayrı bir değişkeni ifade eder. Ayrıca bu değişkenler (rastgele değişkenler, random variables) arasındaki sıralama da bayes ağları ile gösterilebilir [11].
Yapay Sinir Ağları	***Çok Katmanlı Algılayıcı;*** MLP örnekleri sınıflandırmak için geri yayılım kullanan sınıflandırıcı içerir. Bu ağ elle yapılandırılabilir, algoritma ile yaratılabilir veya her ikisi de olabilir. Ağ ayrıca görüntülenebilir ve eğitim zamanı süresince modifiye edilebilir [13]. Çok katmanlı algılayıcı bazı zor ve farklı problemleri başarılı bir şekilde çözmek için uygulanır. Oldukça popüler olan hatanın geriye yayılması mantığına dayalıdır. Bir çok katmanlı ağ üç önemli özelliğe sahiptir: (1) Ağdaki her nöron modeli çıkışta non-lineerlik içerir. Burada önemli bir nokta, non-lineerliğin Rosenblatt'ın perseptronunda kullanılan keskin-geçişli fonksiyona göre yumuşak geçişli olmasıdır. (2) Ağ, çıkış ya da girişe ait olmayan bir ya da daha fazla saklı nörona sahiptir. Bu

	saklı nöronlar ağın karmaşık işleri öğrenmesini sağlar. (3) Ağda, her nöron birbiriyle bağlıdır. Bağlantılardaki bir değişiklik, sinaptik bağlantılarda ve ağırlıklarda değişikliğe neden olur. Geriye yayılım algoritmasının gelişimi, nöron ağlarında bir dönüm noktasıdır [18]. ***Ardışık Minimal Optimizasyon Algoritması (SMO);*** SMO, esas itibariyle destek vektör kullanan bir algoritmadır. Çok terimli kernel kullanarak destek vektör sınıflandırıcıyı eğitmek için SMO Algoritmasını uygular. Bu uygulama global olarak bütün kayıp değerleri yenisiyle değiştirir ve nominal öznitelikleri ikili olanlara dönüştürür. Ayrıca bütün öznitelikleri önceden tanımlanmış değerlerle normalize eder [14].
Karar Ağaçları	***J48;*** J48 algoritması; C4.5 karar ağacı algoritmasının WEKA'ya uyarlanmış versiyonudur. Karar ağacı algoritmaları durumlar veya örnekler kümesiyle başlar ve yeni durumları sınıflandırabilmek için kullanılan ağaç veri yapısı yaratır. Her durum sayısal veya sembolik değer alabilen öznitelikler kümesi ile tanımlanır. Her bir eğitim durumu sınıfın adıyla ifade edilen etikettir. Karar ağacının her bir iç düğümü test içerir, testin sonuçları o düğümden hangi dalın takip edileceğine karar vermek için kullanılır. Yaprak düğümleri testler yerine sınıf etiketleri içerir. Sınıflandırma modu test durumunda, etiket yoksa yaprak düğümüne ulaşır, C4.5 orada bulunan etiketi kullanarak sınıflandırır. Bu özellikler karar ağaçlarını sınıflandırma için değerli ve popüler araçlar yapar [12,15,16].

3. Materyal

Yıldız Teknik Üniversitesi, Bilgisayar mühendisliği bölümü, Sistem Analizi ve Tasarımı dersinde 86 öğrenciye projeleri ile ilgili anket yapılmıştır. Bu derslerdeki projeler bir dönemde yapılmaktadır. Proje gruplarındaki öğrenci sayıları derslere göre değişmekte olup 2 ile 6 kişi arasındadır. Önerilen proje konuları ve grup arkadaşları öğrenciler tarafından belirlenmektedir. Öğrencilere yapılan bu anketler, ders projeleriyle ilgili olup 20 anket sorusu analizde kullanılmıştır. Bu anket sorularının bir kısmı öğrencilerin sosyal yeteneklerini ölçen sorulardan oluşurken bir kısmı da teknik yeteneklerini ölçen sorulardan oluşmaktadır. Anketlerle öğrencilere yöneltilen ve önemli olarak görülen soruların bazıları Tablo 2'de gösterilmiştir. Bu çalışmada veri setlerinde yazı kalabalığı oluşmaması ve değerlendirme aşamalarında kolaylık sağlamak için her veri setindeki sorular "Soru1", "Soru2" ve "Soru3" gibi değerlerle gösterilmiştir. Tablolarda hangi sorunun hangi değerle ifade edildiği de görülebilmektedir.

Tablo-2: Sistem Analizi ve Tasarımı dersi anketinin önemli görülen soruları

Sistem Analizi Ve Tasarımı Soruları	
Soru1	Yazılım geliştirici, analiz ve tasarımdan elde edilen süreç ve raporları ne ölçüde kullanmıştır?
Soru5	Proje içi ekip uyumunuz ne ölçüdeydi?
Soru7	Proje sonunda elde ettiğiniz ürünü nasıl değerlendiriyorsunuz?
Soru10	Yazılıma ait fizibilite analizi yapmak için kullandığınız bir araç ve/veya yöntem var mı?
Soru11	MS project ile belirlediğiniz proje geliştirme zamanı ile sürecin tamamlanmasından sonra elde edilen proje geliştirme zamanı birbirleri ile paralellik göstermekte midir?
Soru14	Veri akış diyagramında tasarladığınız proje ile geliştirdiğiniz projenin varlık, modül ilişkileri ve veri akışları ne ölçüde paraleldir?
Soru15	Yazılım geliştirici, uygulama geliştirme esnasında sistem analiz ve tasarım argümanlarının değişimini gerekli görmüş müdür?
Soru16	Veri tabanı oluşturulurken ER diyagramından yararlandınız mı?

4. Uygulama

Sınıflandırma için NB, BayesNet, MLP, SMO, IBk, KStar ve J48 gibi veri madenciliği algoritmaları analiz edilerek en başarılı algoritma belirlenmiştir. En başarılı algoritma belirlenirken algoritmaların doğruluk ölçütü oldukça basit ve önemli bir kriterdir ancak sadece doğruluk değerine bakılarak en başarılı algoritma seçilmesi sağlıklı olmaz. Algoritmaların başarıları değerlendirilirken Doğruluk değerinin yanı sıra Kappa istatistiği, MAE ve RMSE kriterleri de göz önüne alınmıştır. MAE ve RMSE negatif yönelimlidirler ve düşük değerler daha iyidir. Bu yüzden düşük MAE değerine sahip algoritmalar daha iyi algoritmalar olarak değerlendirilir. Değerlendirmelerimiz anket soruları üzerinden yapıldığı için Kappa değeri de oldukça önemli bir kriterdir. Kappa değeri ne kadar 1'e yakın olursa doğruluk değerinin soruya verilen cevaptaki yığılmaya bağlı olmadığının yani başarının rastgele olmadığını gösterir.

Değerlendirmelerin yapılabilmesi için anket kağıtlarından elde edilen veri seti, dijital ortama aktarılmıştır. Dijital ortama aktarılan anketlerin içeriği Şekil 3'de verilmiştir.

Veri madenciliği algoritmasının uygulanacağı veri kümesinde Şekil 3'de de görüldüğü gibi bazı şıklar hiç işaretlenmemiş, bazı sorulara ise birden fazla cevap verilmiştir. Bu tür kayıp verilere sahip bir veri tabanına uygulanabilecek yöntemlerden ilki kayıp verinin bulunduğu kaydı veri tabanından çıkarmaktır. Eğer kayıp verili kayıt sayısı, toplam kayıt sayısına göre oldukça az ise bu kayıtların veri tabanından çıkarılması mümkündür. Diğer taraftan kayıp veri sayısı yüksekse veya bu kayıtlara ait diğer değerler önemliyse kayıp değerlerin yerine genel bir sabit kullanılabilir ya da kayıp verilerin yerine tüm verilerin ortalama değeri kullanılabilir [17]. Kayıp verilerin yaratacağı sorunları ortadan kaldırmak için ilgili sorulara cevap atanması kayıp değer yerleştirme (Missing Value Replacement) işlemine göre yapılmıştır. Kayıp değer yerleştirme yaklaşımına göre cevabı boş olan soruya cevap olarak ilgili soru

Gıda toptancısı	a	c	a	b	b	b	a	a	c	b	c	e	toplam 7,	b	a	c	b	c	a
Mutfak Robotu	a	c	b	d-a-e-b-c	c	c	b	b	c	a-b	-	-	-	-	-	-	-	-	-
Yıldız-Air Bagaj Otomasyon ve Havayolu	a	c	a	d-e-c-a-b	c	d	b	b	b	b	a	b-e	3-7	b	a	b	b	c	b
YYZ Yayınevi E-ticaret	b	c	b	d-b-e-a-c	b	b	b	b	b	b	b	c	b	c	c	b	b	b	a
Başarım Yazılım	a	c	c	a-e-d-b-c	c	d-e	b	b	b	b	a	c	2-7	b	a	c	a	c	b
Yolcu Bilgi Sistemi	a	c	a	b	b	b-f	b	c	c	b	c	c	4-7	b	a	b	b	b	b
Mutfak Robotu	b	d	b-c	b-c-d-a-e	c	e	b	d	b	b	a	b-c	2-5	b	c	c	b	b	a
Bugün Nereye Gitsem?	a	c	c	a-e-b-e-d ?	a	a-e-f	b	b	b	b	c	a-b-d	toplam 7, 5 sistem	b	a	c	b	c	b
Bankamatik Projesi	a	c	a	a-e-c-d	b	a	a	b	b	b	c	c	3-7	b	a	b	b	b	b
Bugün Nereye Gitsem?	a	c	c	e-a-b-c-d	a	f	b	b	c	a-b	a	c-d	4-5	b	a	b	b	b	b
Bugün Nereye Gitsem?	b	d	a	a-e-b-d-c	b	b	a	c	c	b	a	c	3-7	a	a	c	b	c	b
Ofis Otomasyon Sistemi	a	c	a	d-e-a-c-b	c	e	b	c	c	b	c	b-c	2-4	b	a	c	b	c	b
Kan Bankası	b	a	a-c	a-b-c-e-d	a	f	c	c	a	b	c	e	1-5	b	c	b	b	a	c
Restaurant Bilgi Sistemi	a	c	a	a-e-b-d-c	c	e	a	c	c	b	a	c	2-6	b	a	c	b	c	a
YYZ Yayınevi E-ticaret	a	c	c	b-c-d-a-e	b	e-f	b	a	b	b	c	a	2 sistem	b	a	b	b	b	b
Dersanalerin Ders-ofis	a	c	c	d-e-b-a-c	c	e	b	a	c	b	c	a-b-d	2-4	b	a	c	b	b	b
Spor Kompleksi	a	c	a	c-e-d-b-a	c	e	b	b	c	b	a	b	2-7	b	a	c	b	c	a
Restaurant Bilgi Sistemi	a	c	a	a-b-e-d-c	c	a	a	b	c	b	a	e	1-7	b	a	c	c	c	b
Rent a Car	a	c	a	b-d-a-c-e	c	e	b	b	c	a	d	b-c	4-7	b	a	c	c	c	b

Şekil-3: Örnek anket içeriği

için en çok cevaplanan şıkkın ataması gerçekleştirilmiştir. Veri setine Replace Missing Values modülü uygulandıktan sonra veri seti WEKA'da analizler için kullanıma hazır hale gelmiştir.

Veri madenciliği analizi için kullanılan Weka Programı csv, arff, c4.5 libsvm, xarff gibi formatları desteklemektedir. Bu nedenle veriler uygun formatlara çevrilerek dosyaların okunması sağlanmıştır. Böylelikle Weka'da okunabilir hale gelen dosyalar veri madenciliği işlemlerini uygulamak için hazır hale getirilmiştir (Şekil 4).

Veriler ön işlemden geçirildikten sonra WEKA'da veri madenciliği yöntemlerinden olan sınıflandırma yöntemi kullanılarak sınıflandırma için en başarılı algoritma keşfedilmiştir. Uygulamada ilk önce eğitim evresi vardır, daha sonra üretilen kestirim modelini test etmek için 10 katlı çapraz onaylama kullanılmıştır.

Proje sonunda elde edilen ürünün değerlendirilmesine (başarılı/başarısız) göre sınıflandırma yapıldığında algoritmalarının 10 katlı çapraz onaylama sonuçları Tablo 3'de gösterilmektedir.

Tablo-3: 10 katlı çapraz onaylama istatistikleri

Algoritma	*Doğruluk (%)*	*MAE*	*RMSE*	*Kappa*
NB	94.19	0.0449	0.1727	0.4173
BayesNet	94.19	0.0423	0.1726	0.4173
MLP	96.51	0.0307	0.1536	0.3886
SMO	96.51	0.23	0.2861	0.3886
IBk	94.19	0.0509	0.1804	-0.019
KStar	95.35	0.0368	0.1635	0.312
J48	95.35	0.0497	0.1735	0.0

Veri madenciliğinde bilgiye erişmede farklı metotlar kullanılmaktadır. Bu metotlara ait pek çok algoritma vardır. Bu algoritmalardan hangisinin daha üstün olduğu üzerine pek çok çalışma yapılmış, yapılan bu çalışmalarda farklı sonuçlar elde edilmiştir. Bunun en önemli sebebi, kullanılan veri kaynağına, veri üzerinde yapılan önişleme ve algoritma parametrelerinin seçimine bağlı olmasıdır. Farklı kişiler tarafından, farklı veri kaynakları üzerinde, farklı parametrelerle yapılan çalışmalarda farklı sonuçlar oluşması doğaldır. Algoritmaların Doğruluk

No.	1. soru Nominal	2. soru Nominal	3. soru Nominal	4. soru_Onemli Nominal	4. soru_Onemsiz Nominal	5. soru Nominal	6. soru Nominal	**7. soru** Nominal	8. soru Nominal	9. soru Nominal	10. soru Nominal	11. soru Nominal	13. soru Nominal	14. soru Nominal	15. soru Nominal	16. soru Nominal	17. soru Nominal	18. soru Nominal	19. soru Nominal	20. soru Nominal
1	a	c	c	c	c	c	e	a	b	c	a	c	b	c	c	b	a	b	c	b
2	c	c	c	b	c	c	f	c	c	a	b	a	b	b	b	b	a	a	d	c
3	a	c	a	b	c	b	b	a	a	c	b	c	b	a	c	b	c	a	c	a
4	a	c	b	d	c	c	c	a	b	c	b	a	b	a	c	b	c	b	d	b
5	a	c	a	d	b	c	d	a	b	b	b	a	b	a	b	b	c	b	c	a
6	b	c	b	d	c	b	b	a	b	b	b	b	c	c	b	b	b	a	a	b
7	a	c	c	a	c	c	e	a	b	b	b	a	b	a	c	a	c	b	d	b
8	a	c	a	b	c	b	e	a	c	c	b	c	b	a	b	b	b	b	a	b
9	b	c	a	b	e	c	e	a	d	b	b	a	b	c	c	b	b	a	d	b
10	a	c	c	a	d	a	e	a	b	b	b	c	b	a	c	b	c	b	d	b
11	a	c	a	a	c	b	a	a	b	b	b	c	b	a	b	b	b	b	b	b
12	a	c	c	e	d	a	f	a	b	c	b	a	b	a	b	b	b	b	d	b
13	b	c	a	a	c	b	b	a	c	c	b	a	a	a	c	b	c	b	d	c
14	a	c	a	d	b	c	e	a	c	c	b	c	b	a	c	b	c	b	d	b
15	b	a	a	a	d	a	f	c	c	a	b	c	b	c	b	b	a	c	d	c
16	a	c	a	a	c	c	e	a	c	c	b	a	b	a	c	b	c	a	d	b
17	a	c	c	b	e	b	e	a	a	b	b	c	b	a	b	b	b	b	d	b
18	a	c	c	d	c	c	e	a	a	c	b	c	b	a	c	b	b	b	c	b
19	a	c	a	c	a	c	e	a	b	c	b	a	b	a	c	b	c	a	b	d
20	a	c	a	a	c	c	a	a	b	c	b	a	b	a	c	c	c	b	d	b
21	a	c	a	b	e	c	e	a	b	c	a	d	b	a	c	c	c	b	d	b
22	a	c	a	d	c	c	e	a	b	c	b	a	a	a	c	b	c	a	d	b
23	a	c	a	a	c	b	d	a	d	b	b	c	a	c	b	b	b	a	d	c
24	a	c	a	d	a	c	d	a	b	b	b	a	b	a	c	b	c	a	a	c
25	b	c	a	a	d	c	e	a	b	c	b	a	b	a	c	c	c	b	d	b
26	b	c	c	d	c	c	e	a	b	c	b	a	b	a	c	b	c	b	d	b
27	a	b	b	b	c	b	e	a	d	b	b	a	b	a	b	b	b	c	c	d
28	b	c	c	a	e	b	f	a	d	c	a	a	a	a	b	b	b	b	d	b
29	b	c	a	b	c	b	f	a	d	c	b	c	b	c	b	b	b	a	d	b
30	b	c	a	a	d	c	e	a	d	b	b	c	b	a	b	b	c	b	d	b
31	a	c	a	e	d	b	e	a	d	b	b	a	b	a	c	a	c	b	a	c
32	b	c	a	d	c	c	b	a	a	b	b	a	b	c	b	c	c	a	d	b

Şekil-4: Anketlere verilen cevapların şıklar ile gösterimi

ölçütü oldukça basit ve önemli bir kriterdir. Ancak doğruluk ölçütü tek başına yorumlanırsa değerlendirme yanlış sonuçlara götürebilir. Bu ölçütü Kappa, MAE ve RMSE ölçütleriyle beraber ele almak gerekir. Bu ölçütler çerçevesinde değerlendirildiğinde, sınıflandırmada en başarılı algoritma Doğruluk, MAE, RMSE ve Kappa değerleri sırasıyla %96.51, 0.0307, 0.1536, 0.3886 olan MLP algoritmasıdır. MAE ve RMSE sonuçları oldukça küçük olup Kappa değeri ise oldukça gelecek vadediyor.

Aynı veri setine veri madenciliğinde, birliktelik kuralı çıkarım algoritmalarından biri olan Apriori algoritması kullanılarak sorularda yer alan yazılım geliştirme süreçlerine ait kavramlar arasındaki birliktelikler belirlenmeye çalışmıştır. Birliktelik kuralında güven değeri çıkarımların gücünü verirken, destek değeri ise kurallardaki örneklerin frekansını vermektedir. Yüksek güven ve güçlü destek değerleri kullanılarak birbirleriyle bağlantılı olan yüksek ilişkili parçaların birliktelikleri çıkartılabilir. Bu nedenle güven ve destek değerlerinin önemli birliktelikleri vermesi için bu iki değerin de belirli eşik değerlerinden yüksek olması beklenmektedir.

Yüzde 70 destek değeri ve yüzde 90 güven değeri için Şekil 5'de görülen kurallar elde edilmiştir.

```
Minimum support: 0.7 (60 instances)
Minimum metric <confidence>: 0.9
Number of cycles performed: 6
Best rules found:

 1. 14. soru=a 61 ==> 7. soru=a 61    conf:(1)
 2. 2. soru=c 78 ==> 7. soru=a 76    conf:(0.97)
 3. 2. soru=c 13. soru=b 64 ==> 7. soru=a 62    conf:(0.97)
 4. 2. soru=c 16. soru=b 64 ==> 7. soru=a 62    conf:(0.97)
 5. 16. soru=b 69 ==> 7. soru=a 66    conf:(0.96)
 6. 10. soru=b 68 ==> 7. soru=a 65    conf:(0.96)
 7. 13. soru=b 72 ==> 7. soru=a 68    conf:(0.94)
 8. 7. soru=a 16. soru=b 66 ==> 2. soru=c 62    conf:(0.94)
 9. 16. soru=b 69 ==> 2. soru=c 64    conf:(0.93)
10. 7. soru=a 82 ==> 2. soru=c 76    conf:(0.93)
```

Şekil-5: %70 Destek ve %90 Güven değeri için elde edilen kurallar

Şekil 5 'de Apriori yaklaşımıyla sorular arasındaki ilişkiler güven değeri sırasına göre sıralanmış bir şekilde görünmektedir. Kurallardan bazıları incelendiğinde şu sonuçlar ortaya çıkmaktadır. Örneğin, 1. kural, ders kapsamında kullanılması beklenen ER Diyagram, Veri Akış Diyagramı, Yapı diyagramını uygulama süresince kullanan öğrencilerin yüzde 100'ü projelerini başarılı bulacaktır. 2. kural, süreç başında tasarlanan yazılım ile süreç sonunda oluşturulan yazılımın birbiriyle %50-75 arasında örtüşmesi durumunda öğrencilerin yüzde 97'si projelerini başaralı bulacaktır. 6. Kural, fizibilite analizi için MS Project kullanan öğrencilerin yüzde 96'sı projelerini başarılı bulacaktır.

5. Sonuçlar ve Öneriler

Yapılan çalışmada Yıldız Teknik üniversitesi, Bilgisayar Mühendisliği bölümü, Sistem Analizi ve Tasarımı dersini alan öğrencilere yaptıkları projeler ile ilgili anket uygulanmıştır. Çalışmadaki hedef proje tabanlı gerçekleştirilen derslerde başarılı projelerin oranını arttırmak ve öğrenciyi yönlendirmektir. Çünkü sektörde yapılan projeler büyük oranda başarısız projelerdir. Başarısız projeler oldukça mali kayıplara neden olmaktadır ve bu başarısız projelerin sebebi akademide alınan eğitimden kaynaklı olduğu düşünülmektedir. Bu nedenle öğrenciler mezun olmadan ve önemli tasarım ve uygulama sorumlulukları almadan önce mutlaka bazı pratik deneyimleri kazandırmak gerekmektedir [1]. Başarısız projelerin oranını düşürebilmek için öğrencilere deneyim kazandırmak gerekmektedir. Bu deneyim de ders kapsamında gerçekleştirilen projelerden kazandırılabilir.

Bu çalışmada öğrenci projelerini sınıflandırmak için NB, BayesNet, MLP, SMO, IBk, KStar ve J48 gibi veri madenciliği algoritmaları analiz edilerek ve en başarılı algoritma belirlenmiştir. Algoritmaların ürettikleri sonuçlar karşılaştırıldığında MLP algoritmasının diğer algoritmalardan daha başarılı olduğunu görülmüştür. MLP algoritması MAE=0.1536, RMSE=0.1536, Kappa=0,3886 Doğruluk=96.51 değerleriyle üstün performans göstermiştir.

Ayrıca birliktelik kuralı çıkarım algoritmalarından biri olan Apriori algoritması kullanılarak sorularda yer alan yazılım geliştirme süreçlerine ait kavramlar arasındaki birliktelikler belirlenmiş olup bu kurallar Bölüm 4'de sunulmuştur.

Bu çalışmada, modellerin oluşturulması için ücretsiz bir yazılım olan Weka aracı kullanılmıştır. Var olan diğer veri madenciliği araçları üzerinde aynı algoritmalar çalıştırılarak sonuçlar karşılaştırılabilir. Anketler daha fazla öğrenciye uygulanarak veriler arttırılabilir. Ayrıca akademi ile sektörün paralel

çalışmasıyla elde edilen öğrenci projeleri üzerinde yapılan analizler karşılaştırılabilir.

6. Kaynakça

[1] Saiedian, H., Bagert, D., Mead, N. *Software Engineering Programs: Dispelling the Myths and Misconceptions*, IEEE Software, 2002, pp. 35-46.

[2] Hilburn, T., Humphrey, W.*The Impending Changes in Software Education*, IEEE Software, 2002, pp. 22-24.

[3] Kurt, Ç., Erdem, O.A.. *Öğrenci Başarısını Etkileyen Faktörlerin Veri Madenciliği Yöntemleriyle İncelenmesi*, Politeknik Dergisi, 2012, pp. 111-116.

[4] Bırtıl, F. S. *Kız meslek lisesi öğrencilerinin akademik başarısızlık nedenlerinin veri madenciliği tekniği ile analizi*, Yüksek Lisans Tezi, Afyon Kocatepe Üniversitesi, Fen Bilimleri Enstitüsü, 2011.

[5] Bozkır, A.S., Sezer, E., Gök, B. *Öğrenci Seçme Sınavında (ÖSS) Öğrenci Başarımını Etkileyen Faktörlerin Veri Madenciliği Yöntemleriyle Tespiti*, 5. Uluslararası İleri Teknolojiler Sempozyumu (IATS'09), Karabük Üniversitesi, 2009, pp. 37-43.

[6] Aydın, S. Veri madenciliği ve Anadolu üniversitesi uzaktan eğitim sisteminde bir uygulama, Doktora Tezi, Anadolu Üniversitesi, Sosyal Bilimler Enstitüsü, 2007.

[7] Ayık, Y.Z., Özdemir, A., Yavuz, U. *Lise Türü ve Lise Mezuniyet Başarısının Kazanılan Fakülte ile İlişkisinin Veri Madenciliği Tekniği ile Analizi*, Sosyal Bilimler Enstitüsü Dergisi, 10(2), 2007.

[8] Karabatak, M., İnce, M. C. *Apriori Algoritması ile Öğrenci Başarısı Analizi*, Eleco' 2004 ElektrikElektronik ve Bilgisayar Mühendisleri Sempozyumu, Bursa, 2004.

[9] Aha, D., Kibler, D. *Instance-based learning algorithms*, Machine Learning, 1991, pp. 37-66.

[10] Cleary, J.G., Trigg, L.E. *K*: An Instance-based Learner Using an Entropic Distance Measure*, Proceedings Tweleveth International Conference on Machine Learning, Tahoe City, California, 1995, pp. 108-114.

[11] Internet: Bayes Ağları (Bayes Network), http://bilgisayarkavramlari.sadievrenseker.com/2008/12/21/bayes-aglari-bayesian-network/, Ekim 2015.

[12] Tan, P.N., Steinbach, M., Kumar, V. *Introduction to Data Mining*, USA, 2006.

[13] Internet: Data Mining Software in Java, www.cs.waikato.ac.nz/~ml/weka/, Mart 2013.

[14] Ardıl, E., *Esnek Hesaplama Yaklaşımı İle Yazılım Hata Kestirimi*, Yüksek Lisans Tezi, Trakya Üniversitesi, Fen Bilimleri Enstitüsü, 2009.

[15] Weı, C., Chıu, T., *Turning telecommunications call details to churn prediction : a data mining approach*, Expert Systems with Applications, 2002, pp. 103-102.

[16] Quinlan, J.R., *Induction of Decision Trees*, Machine Learning, 1986, pp. 81-106.

[17] Silahtaroğlu, G., *Kavram ve Algoritmalarıyla Temel Veri Madenciliği*, Papatya Yayıncılık, İstanbul, Türkiye, 2008.

[18] Kıyan, T., Yıldırım, T., *Eğiticili Ve Eğiticisiz Nöral Algoritmalar Kullanarak Göğüs Kanseri Teşhisi*, Elektrik -Elektronik-Bilgisayar Mühendisliği 10. Ulusal Kongresi, İstanbul, Türkiye, 2003.