WEB İSTATİSTİKLERİNDE MAKİNE ÖĞRENMESİ ALGORİTMALARI İLE KRİTİK PARAMETRE TESPİTİ

Harun BAYER[1], Tevfik ÇOBAN[2]

Özet

Web sitelerinin internet ziyaretçileri tarafındaki yansıması ile oluşan web sitesi istatistiklerinin bu sektördeki önemi artmaktadır. Web istatistikleri, web sunucuları tarafında tutulabilen kayıtlardır ve web sitelerinin popülaritesini kıyaslamada net ve doğru bir bilgi sunmaktadırlar. Oluşan bu değerler bütününün ağ trafiğinde anlamlandırılabilmesi için sınıflandırma veya kümeleme işlemlerine tabi tutulması gerekmektedir. Bu bağlamda, Türkiye'de yaygın olarak tıklanan web sitelerinin istatistiksel verileri kullanılarak, makine öğrenmesi algoritmaları ile analizi bu çalışmanın amacı olmuştur. Elde edilen gerçek veriler üzerinde makine öğrenmesinin nasıl gerçekleştiği ve veriler arasında web site trafiğindeki en belirleyici parametreler tespit edilmiştir. Bu tespitler, gözetimli ve gözetimsiz öğrenme algoritmaları ile gerçekleştirilmiştir. Eğitim, test ve çapraz doğrulama gibi farklı seçeneklerle ayrıntılı olarak incelenen bu algoritmaların birbirine göre başarı ve performans kıyaslaması yapılarak web siteleri parametre analizleri araştırılmıştır.

Anahtar Kelimeler: Makine öğrenmesi, parametre, sınıflandırma, web istatistik

DETERMINATION OF CRITICAL PARAMETERS THROUGH MACHINE LEARNING ALGORITHMS IN WEB STATISTICS

Abstract

The importance of website statistics which consists of reflections on the internet visitors side is increasing in this sector. Web statistics can offer clear and accurate information in comparing the popularity of websites as the records are kept by web servers. This entirety of values must be subjected to sorting or clustering processes to be justified in network traffic. In this context, the purpose of this study is to analyze the machine learning algorithms with using statistical data of common websites clicked in Turkey. The process of actual machine learning and the most determinant parameter in website traffic among data are determined through the obtained actual data. These findings are obtained from the supervised and unsupervised learning algorithms. The websites parameters' analysis is made through the comparison of the success and performance of these algorithms. On the other hand, these algorithms are investigated in detail with various options such as training, testing, and cross-validation.

Keywords: Machine learning, parameter, classification, web statistic

[1] Öğr. Gör., Kırklareli Üniversitesi, harunbayer@gmail.com

[2] Mühendis, Adli Tıp Kurumu, tevfikcoban@gmail.com

Giriş

Kurumların imajını güçlendirmek ve sektördeki pazar payını artırmak, e-ticaret yoluyla satış yapmak, dosya paylaşımı, arkadaşlık siteleri gibi sitelerle ücretli veya ücretsiz üyeleri artırmak, siteye alınan reklamlardan gelirleri artırmak web sitesi sahiplerinin temel amaçları arasında yer alır. Bu nedenlerle web sitelerinin popülerliliği, sahiplerine avantaj oluşturan bir husustur. Bunun yanı sıra reklam veren, bir web sitesine reklam vermek istediğinde hangi siteyi neye göre değerlendirmelidir? Reklam verilen bir web sitesinin kazandıracağı müşteri potansiyeli sadece ziyaretçi sayısına mı bağlıdır, yoksa o web sitesinde çok zaman geçirilmesine mi? Web istatistikleri değerleri bakımından iyi ya da kötü olan bir web sitesinin bu konumunu en çok belirleyici istatistikî parametre hangisidir? Maddi gelir ya da hobi amaçlı olarak bir kişi, çok ziyaret edilen bir web sitesi isteği varsa, hangi konuya yönelik türde bir web sitesi hazırlamalıdır? vb. sorular web siteleri istatistikleri tutulması, analiz edilmesi, değerlendirilmesi ve yorumlanması sürecini başlatan sorulardan birkaçıdır.

Çalışma konusu olan diğer kavram makine öğrenmesi; örnek verileri kullanarak, ya da geçmişteki deneyimlerden yararlanarak, tümevarım yöntemiyle tanımlayıcı ya da tahmini çıkarımlar yapacak şekilde bilgisayar programlamak olarak tanımlanmaktadır (Alpaydın, 2004). Diğer tanımıyla makine öğrenmesi, veritabanlarına dayalı veriler arasındaki gizli kalmış karmaşık örüntünün tespit edilmesi ve anlamlı desen çıkarımı için istatistik ve bilgisayarın hesaplama gücünden yararlanır.

Web istatistiklerinin değerlendirilmesi ve internet ağ trafiğinin yönetimi konusundaki genel beklentiler ışığında, yapılan bu çalışmada Türkiye ağında bulunan en çok ziyaretçi almış 1000 adet web sitesinin web istatistikleri bir Google servisi olan '*DoubleClick Ad Planner*' aracılığı ile elde edilmiştir. Elde edilen veriler ön işlem sürecinden geçirilerek, yapısal veri haline dönüştürülmüştür. Daha sonra verilerin tümü, makine öğrenmesi algoritmalarını bünyesinde barındıran bir yazılım olan WEKA programına yerleştirilerek hem hedef niteliklerin hem de algoritmaların birbirleri ile farklı performans ölçütleri kapsamında başarı kıyaslaması yapılmıştır. Böylece web sitelerini sınıflandırmada en kritik parametre tespit edilmiş olup aynı zamanda hangi algoritmanın daha başarılı olduğu saptanabilmiştir.

Yöntem

Makine Öğrenmesi Yöntemleri

Bilgisayar sistemlerinde, bilgisayarları programlamak çeşitli problemlere her zaman çözüm getirmeyebilir. Çünkü girdi verilerine karşılık istenilen çıkışı elde etmenin bilinen bir yöntemi olmayabilir. Gerekli yöntem açıkça ifade edilemediğinden, bu örnekler klasik programlama yaklaşımı yerine deneyimlerden yararlanarak çözülmeye çalışılmaktadır. Bilgisayarın öğrenmesi de giriş/çıkış işlevselliğini örneklerden öğrenmesidir. Bu bağlamda programları sentezlemek için örnek kullanma yaklaşımına makine öğrenme yöntemi denir (Cristiannive ark. 2000).

Gözetimli Öğrenme (Supervised Learning)

Makine öğrenmesi yöntemlerinden olan gözetimli öğrenmede, işlem eğitilmiş veriler üzerinden gerçekleştirilir. Bu yöntemde girdi ve çıktı önceden belirlenmektedir.

Bu girdi ve çıktı verileri arasındaki ilişki ile makinenin işleyişi öğrenmesi gerçekleştirilmektedir. Bu öğrenme yapısına göre de makine yeni girdi verilerine karşı çıktı olabilecek tahminler yaparak, sınıflandırma gerçekleştirebilmektedir(Uzun. 2007).

Bayes Sınıflandırma Algoritmaları

Olasılık kuramı içerisinde yer alan Bayes teoremi, rastlantısal değişkenler için olasılık dağılımı içinde şartlı olasılıklar ile aykırı olasılıklar arasındaki etkileşimi gösterir. Bayes teoremi, hedef değişkenle bağımsız veriler arasındaki etkileşimi analiz eder. Bayes teoremi tahmin edici ve tanımlayıcı bir sınıflama algoritmasıdır (Amasyalı ve Bilgin, 2015).

NaiveBayes Algoritması

NaiveBayes yüksek boyutlu metin verilerinin analizinde, hesaplama ve sınıflandırma doğruluğu nedeniyle makine öğrenmesinin popüler bir algoritmasıdır. NaiveBayes sınıflandırma algoritmasında, girdi verisinde birbirinden bağımsız özellikleri olan her bir parametrenin sonuca olan etkisi olasılıksal olarak hesaplanmaktadır. Bundan dolayı, veri sınıflandırması aslında sınırları koşullu olasılık değerleri ile belirli olan bir tahmindir (Lee.2015).

F=$\{F_1, F_2, \ldots, F_m\}$ birbirinden bağımsız m adet niteliğe ait sınıf değerlerinden oluşan veri örneği ve $C_1, C_2, \ldots, C_n$bağımlı sınıf değişkeni olarak düşünülürse, Bayes teoremi yardımıyla aşağıdaki denklem elde edilir:

$$P(C_i|F) = \frac{P(F|C_i)\,P(C_i)}{P(F)} \qquad (1)$$

Burada her $P(C_i|F)$ değeri için $P(F)$ sabit kalacağından ve sınıf seçiminde maksimum benzerlik dikkate alınarak büyüklük karşılaştırması yapılacağından dolayı paydayı, yani $P(F)$ değerini, hesaplara katmaya gerek kalmamaktadır (Özkan, 2008). Çünkü F verisi sabit bir veri olup, sınıfı belirlenmesi istenen F verisi için (1)' deki pay değerinde en büyük değeri sağlayan C_i sınıfı seçilerek, veri örneği sınıfı belirlenmiş olur. F verilerine ait her C_i sınıfı için hesaplanan bu değerlerden büyük olan i sınıfı seçileceğinden dolayı Naive Bayes algoritması için sonuç ifadesi;

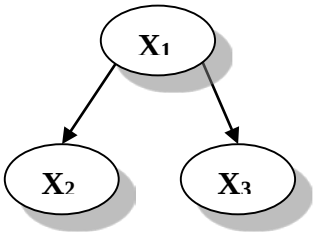
$$\arg \max_c \{P(F|C_i)P(C_i)\} \qquad (2)$$

olur. Bu ifade "maximum a posterioriclassification (MAP)" olarak bilinir(Cristiannive ark. 2000).

Bayes Ağı Algoritması

Bayes ağı algoritmasında, çeşitli olaylarda değişkenler arası ilişkileri belirlemek amacıyla grafiksel modellerden yararlanılmaktadır (Dünder ve ark. 2013). Bayes ağları uygulamalarında her bir değişken düğümlerle gösterilir. Değişkenler arasındaki etkileşimi, şartlı olasılıklarla ve grafik olarak temsil ederler. Bu grafik, kullanıcı ara yüzü olarak kullanılır ve modellenen problemin değişkenleri arasındaki etkileşimi görsel olarak yansıtır. Bayes ağları nedeni belli olmayan karmaşık problemler üzerine anlamlı sonuçlar üretmek için geliştirilmiştir. Eldeki problemin doğası gereği ya da verilerin eksik

olması nedeniyle net bir sonucun elde edilmediği durumlarda başarılı bir performans göstermektedir (Sorias, 2014).



Şekil 1. Örnek bir Bayes ağı

Şekil 1' de verilen basit bir Bayes ağının birleşik olasılık dağılımı (3)'de verilmektedir:

$$P(X_1, X_2, X_3) = P(X_2|X_1).P(X_3|X_1).P(X_1) \quad (3)$$

Şekil 1 'de X_1'den X_2'ye bir kenarın var olduğu görülmektedir. Bu durumda X_1, X_2'nin ebeveyni olur ve X_1'in X_2 üzerinde doğrudan bir etkisi vardır. X_2'nin bir ebeveyni olduğundan, kendisinin yerel olasılık dağılımı da koşullu olur. i=1, 2, 3,…..n olmak üzere her X_i'ye ait ebeveynler kümesi ebeveyn(X_i) olarak gösterilirse Bayes ağları için genelleştirilebilecek bileşik olasılık dağılımı ifadesi (4)'deki hali alır (Özkan, 2008).

$$P(X_1 = x_1, \dots, X_n = x_n) = \prod_{i=1}^{n} P\left(X_i = x_i \,\middle|\, \text{ebeveyn}(X_i)\right) \quad (4)$$

Destek Vektör Makineleri Algoritması

Sınıflandırma algoritmaları arasında popülerliği yüksek, yeni bir sınıflama yöntemi olan destek vektör makineleri (SVM) ilginç teorik ve pratik özellikleri nedeniyle birçok alanda kullanılmaktadır. İstatistiksel öğrenme kuramına dayalı olan SVM, genel bir sınıflandırma yöntemidir (Habibi ve ark. 2015).

Destek Vektör Makineleri, doğrusal olmayan örnek uzayını, örneklerin doğrusal olarak ayrılabileceği bir yüksek boyuta aktararak, farklı örnekler arasındaki maksimum sınırın bulunması esasına dayanır. Bu alanda karşılaşılan problemlerin büyük çoğunluğu, birçok farklı bileşenden oluşan problemlerdir ve doğrusal olarak ayrılmış bir yapı halinde değildirler (Güneş ve Yiğit, 2012). Doğrusal olarak ayrılmış olan veriler arasında maksimum sınırın direkt olarak bulunması basit olmasına rağmen, doğrusal olarak ayrılamayan veriler öncelikle doğrusal olarak ayrılabilecekleri farklı bir uzaya aktarılmaktadır (Silahtaroğlu, 2008). Doğrusal olarak ayrılabilme durumunda $x_i \epsilon R^d$özellikler vektörü ve her biri y_i={-1,+1} ile gösterilen sınıflardan birine ait olmak üzere n adet eğitim verisi $\{x_i, y_i\}$ şeklinde ifade edilsin. Bu durumda, ayırıcı aşırı düzlem için fonksiyon aşağıdaki gibidir:

$$f(x) = w^T.x + b = \sum_{i=1}^{n} w_i x_i + b \quad (5)$$

Burada w, aşırı düzlemin normalini, b sabiti ise sapma değerini ifade eder. |b|/||w||

aşırı düzlemin orijine olan uzaklığı, x aşırı düzlem üzerinde olan herhangi bir nokta olmak üzere aşırı düzlem üzerindeki noktalar cinsinden $f(x) = w^T.x + b = 0$ koşulunu sağlamaktadır. Dolayısı ile ayırıcı aşırı düzlem verileri iki sınıfa böleceğinden (6)'da ki gibi yazılır.

$$\text{Sınıf 1}: \quad f(x) > 0 \text{ için } y_i = +1;$$

$$\text{Sınıf 2}: \quad f(x) < 0 \text{ için } y_i = -1 \tag{6}$$

K En Yakın Komşu Algoritması

K en yakın komşu algoritması, öznitelik uzayındaki eğitim örneklerine dayanarak nesneleri sınıflandırmak için kullanılan, ileriye dönük istatistiksel sınıflandırma algoritmalarından biridir. Sınıflandırılması istenen bir verinin en yakın k komşuluktaki verileri temel alarak sınıflandırıldığı bir yöntemdir. Bu algoritmada sınıfı bilinmeyen bir verinin, k en yakın komşulara olan benzerlikleri hesaplanır ve sınıflara ataması gerçekleştirilir. K en yakın komşu algoritmasında eğitim aşaması çok hızlıdır fakat test kısmı hem zaman hem de bellek açısından maliyetlidir (Saini ve ark., 2013).

Karar Ağaçları İle Sınıflandırma Algoritmaları

Karar ağaçları, eğitim ve testinin hızlı olması, sonuçlarının daha kolay yorumlanabilmesi sebebiyle çok sık kullanılan sınıflandırma yöntemlerinden biridir (Kaya ve Yıldız, 2014). Bir veri kümesi için, aynı niteliklerle çok farklı şekillerde karar ağaçları kurulması mümkündür. İdeal olan karar ağaçlarının kurulmasında dikkat edilmesi gereken en önemli aşama en ayırıcı niteliklerin tespitidir. Çünkü karar ağaçlarında temel amaç, bir veri kümesini niteliklerine göre sorgulamalar yaptırarak en hızlı şekilde sonuca ulaşabilmektedirler.

ID3 Algoritması

ID3 algoritmasında, hedef kümenin en ayırıcı niteliklerini belirlemek üzere entropi kurallarını temel alan bilgi teorisi kullanılmaktadır. Entropi, bir sistemdeki belirsizlik olarak tanımlanabilir. Tek hedef nitelikli karar ağaçlarında ID3 algoritması, bilgi kazancı yaklaşımını kullanmaktadır (Kavzoğlu ve Çölkesen, 2010).

C4.5 Algoritması

ID3 algoritması, karar ağaçları oluşturmada kullanılmasına rağmen bazı noktalarda yetersiz kalan bir algoritmadır. Bu eksiklerin giderilmesi için C4.5 algoritması geliştirilmiştir. ID3 algoritmasının yetersiz olduğu noktalardan birisi niteliklerin, kategorik veriler yerine, birbirleri ile çoğunlukla eşit olmayan sayısal değerlerden oluşmasında görülmektedir (Yıldırım, 2003). Bu durumda çok sayıdaki sayısal değerlerin her biri için bir dal oluşturulacağı için karar ağacının gereksiz yere çok büyümesine ve dallanmasına yol açılır. Bu da sınıflandırmayı çok kullanışsız hale getirmektedir.C4.5 algoritması, bu soruna çözüm olarak söz konusu niteliğe ait verilerde en büyük bilgi kazancını sağlayacak eşik değer yaklaşımını getirmektedir (Özkan, 2008). Sürekli değişkenler içerisinde, uygun eşik değeri bulunduktan sonra ikili ya da daha çok bölünme ile veri kümesi bölünebilir. Eşik değer, küçükten büyüğe doğru sıralanan verilerin ortancası olarak alınabilir. Böylece sayısal verilerden oluşan bu küme, eşik değerden

küçük ve büyük veriler diye iki alt kümeye ayrılmaktadır. C4.5 algoritmasının ID3 algoritmasına göre başka bir üstünlüğü ise kayıp verilerle de çalışabilmesidir. Algoritma, bu durumda eksik veriye bağlı diğer tüm değerleri göz önüne almadan entropi ve kazanç değerlerini hesaplar. Ancak bulunan kazanç değerleri bir düzeltme faktörü yardımıyla yeniden hesaplanmaktadır (Özkan, 2008).

Gözetimsiz Öğrenme (Unsupervisedlearning)

Gözetimsiz öğrenme, girdi verilerinin en uygun gösterim şeklini belirleyen, bir girdi kümesi modellemesinde kullanılan öğrenme türüdür. Bu öğrenmede, girdi verilerine karşılık gelen herhangi bir çıktı verisi (sınıf ya da etiket bilgileri) yoktur. Hiçbir girdinin hangi sınıfta olduğu bilinmemektedir. Sistem, sınıfları bilinmeyen girdi kümesini parçalara ayırarak girdi verilerinin gruplandırılmasını ya da kümelenmesini sağlamaktadır. Bu nedenle kümeleme, en temel gözetimsiz öğrenme yöntemidir (öztemel,2006).

K –Means Algoritması

K-means algoritması kolay uygulanabilir, uygulama alanı geniş bir veri demetleme algoritma olmayı sebebiyle en sık kullanılan gözetimsiz algoritmalarındandır (Çetin ve Hacıömeroğlu, 2013).

K-means algoritması, elde mevcut bulunan verileri, kullanıcı tarafından belirlenen k parametresi kadar kümeye ayıran, gerçeklemesi kolay gözetimsiz öğrenme algoritmalarından biridir. Bu algoritma, benzer özellik gösteren verilerin, bir arada kümelenmesi esasına dayanır. Algoritmadaki amaç, oluşturulan k adet kümenin, kendi içlerinde benzerliklerinin maksimum, birbirleri arasındaki benzerliklerinin ise minimum olmasını sağlamaktır (Bülbül ve ark. 2009).

Hiyerarşik Kümeleme Algoritmaları

Hiyerarşik kümeleme algoritmaları, veriler arasındaki uzaklık bilgilerinden yararlanarak birleşme ya da bölünme kurallarının çıkarılmasını sağlayan algoritmalardır. Algoritma, başlangıçta bir veritabanındaki verilerin her birini bir küme olarak kabul eder ve aşama aşama birleştirerek belirli bir sıra ile tek bir küme elde edilmesini (birleştirici hiyerarşik kümeleme) sağlar. Bu işlemin tersi olarak da, verileri tek bir küme kabul ederek, yine belirli bir sıra ile bölünebilmesini (ayrıştırıcı hiyerarşik kümeleme) sağlar. Buradaki önemli nokta, hangi verinin hangi kümeye kaçıncı sırada dâhil edildiğinin ya da ayrıldığının bulunmasıdır (Bülbül ve ark. 2009).

Çalışmada Kullanılan Verilerin Elde Edilmesi ve Veri Dönüşümü

Reklam verenlerin, reklam vermek istedikleri siteleri bulmalarını sağlayan web tabanlı bir yazılım olan ‘Doubleclick Ad Planner’ Reklam verenler bu yazılım aracılığıyla kategorisine, ziyaretçisine ve farklı birçok niteliğe bakarak site seçimi yapabilir http://www.damlakaraman.com.tr/google-doubleclick-nedir, (2013)

Web tabanlı çalışan DoubleClick Ad Planner’da ‘Araştırma’ üst menüsü yardımıyla çeşitli filtreleri kullanarak belirli düzeyde web sitesi verilerine

ulaşılabilmektedir. Bu menü aracılığı ile 'Kitleye göre arama' ara yüzü ile web sitelerine ait verilerin olduğu listeye ulaşılmış ve bu veriler kaydedilmiştir. Listede yer alan 1000 web sitesi için ayrı ayrı site profili incelenerek gerekli veriler için kayıt tutularak veritabanı oluşturulmuştur.

Türkiye'deki ziyaretçi sayısı en fazla olan ilk 1000 siteye ait 6 adet farklı niteliğe ait veriler DoubleClick Ad Planner ile elde edilmiştir. Bu nitelikler Tablo 1.'de verilmektedir.

Tablo 1. Google Ad Planner İle Elde Edilen Web Sitesi Verilerine Ait Nitelikler

	Web sitesi adı
1	Site kategorisi
2	Erişim
3	Farklı ziyaretçiler (tahmini çerezler) sayısı
4	Sayfa görüntüleme sayısı
5	Sitede geçirilen ortalama süre
6	Ortalama ziyaret sayısı

Site kategorisi, web sitesinin yönelik olduğu konu ya da verdiği hizmetlere dönük olarak sitenin içeriğini yansıtan niteliktir. Belirli bir ay boyunca, belirlenen ülke için tahmini toplam internet kullanıcısı yüzdesini ifade eder. Farklı ziyaretçiler (tahmini çerezler), sitedeki DoubleClick Ad Planner'ın algoritmaları tarafından belirlenen yaklaşık çerez sayısıdır (http://www.turkeyforum.com/satforum/archive/index.php/t-635294.html). Farklı ziyaretçiler (tahmini çerezler) değeri, gerçek farklı ziyaretçi sayısı değildir ancak web sitesi tıklama analizlerinde kıyaslama yapılarak kullanılacağından yararlı olabilecek bir parametredir. Gerçek farklı ziyaretçi değeri olmamasının birkaç sebebi vardır: Çünkü çerezler kendiliğinden geçici süreli olabilir veya kullanıcı çerezi kendisi silebilir. Bu durumda aynı IP adresinden web sitesine giren bir kullanıcı tekrar çerez almış olur. Ya da kullanıcılar, birden fazla web tarayıcısı kullanıyorsa her web tarayıcısı için ayrı ayrı çerez alabilir. Sayfa görüntüleme sayısı, tüm kullanıcıların belirli bir ay boyunca, bir sitedeki sayfaları toplam görüntüleme sayısıdır. Sitede geçirilen ortalama süre, ortalama olarak her ziyaretçinin sitede geçirdiği süredir. Ortalama ziyaret sayısı ise, web sitesi ziyaretçileri başına düşen ziyaret sayısını ifade etmektedir (Demirci, 2007).

6 adet niteliğe ilişkin sürekli verilerdeki değişimler ve kategorik verilerdeki benzerlikler incelenerek dönüştürme işlemleri yapılmıştır.

Site kategorisi: DoubleClick Ad Planner ile elde edilen 1000 web sitesine ait toplamda 200'ün üzerinde farklı site kategorisi bulunmaktadır. Hem sınıflandırma problemlerinin yaşanmaması açısından hem de birbiri ile çok yakın ve ilişkili kategorilerin farklı bir sınıf gibi davranmasını engellemek amacı ile gruplanarak 20 ana web sitesi türü elde edilmiştir. Bu site türleri ve temsil edilen kategoriler Tablo 2'de verilmektedir.

Tablo 2. Web Sitesi Kategorilerinin Veri Dönüşümü

	Veri dönüşümünden önce	Veri dönüşümünden sonra	
No	Site Kategorisi (*)	Web Sitesi Türü	Veri Sayısı
1	Bilim Kurumları, Devlet, Devlet Arşivleri, Hukuk, Hukuk ve Devlet Hizmetleri, Kamu Güvenliği, Kamu Maliyesi, Yönetim, Yüksekokullar Eyalet Yönetimi ve Yerel Yönetim	Resmi Kurum	43
2	Web Portalları , Arama Motorları (**gerçekte arama motoru da içeren bazı internet portalları, Arama Motoru kategorisi adıyla kayıtlı olduklarından İnternet Portalına dahil edilmiştir.*)	İnternet Portalı	26
3	Açık artırmalar, Alışveriş, Alışveriş Portalları, Bilet Satışları, DVD ve Video Alışverişi, Fiyat Karşılaştırmaları, İthalat ve İhracat, Teklifleri, Pazarlama Hizmetleri	e-Ticaret	71
4	Araç Alım-Satım, Emlak, Gayrimenkul İlanları, Konut ve Arazi Geliştirme, Otomobiller ve Araçlar,Seri İlanlar, İş İlanları,İstihdam ve Personel Alımı,İşletme İlanları ve Kişisel İlanlar,İnsan Kaynakları	İlanlar	32
5	Basit Oyunlar, Bilgisayar ve Video Oyunları, Çevrimiçi Oyunlar, Devasa Çok Oyunculu Oyunlar, Kağıt Oyunları, Nişan Oyunları, Oyunlar, Spor Oyunları	Oyun	90
6	Ağ Güvenliği, Ağ Oluşturma, Bilgisayar Güvenliği, Donanım, İnternet ve Telekom, Mobil Uygulamalar ve Eklentiler, Mühendislik ve Teknoloji, Proxy ve Filtre Kullanımı, Teknoloji Haberleri, Windows OS	Bilişim - Teknoloji	30
7	Eğitim, İlk ve Ortaöğretim (K-12), Okul Öncesi Eğitim, Test Hazırlığı	Eğitim	30
8	Demografi, Kişi Arama, Sosyal Ağlar, Blog Kaynakları ve Hizmetleri	Sosyal Paylaşım	24
9	Amerikan Futbolu, Basketbol, Spor ve maç sonuçları, Futbol, Spor, Spor Haberleri	Spor	29
10	Bankacılık, Emtia ve Vadeli İşlemler Ticareti, Finans, Kredi Kartları, Kredi ve Borç, Muhasebe ve Teftiş, Para Birimleri ve Döviz, Ticaret Hizmetleri ve Ödeme Sistemleri	Ekonomi - Bankacılık	32
11	Forum ve Sohbet Sağlayıcıları	Forumlar	38
12	Dosya Paylaşımı ve Barındırma, Fotoğraf ve Görüntü Paylaşımı, Fotoğraf ve Video Yazılımları, İnternet İstemcileri ve Tarayıcıları, İnternet Yazılımları, Ortam Yürütücüleri, Ücretsiz Çevrimiçi Öğeler	Dosya Paylaşımı	71
13	Cinsel Eğitim ve Danışmanlık, Ereksiyon Sorunları, Kilo Verme, Sağlık, Sağlık Haberleri, Sağlık Sigortası, Sağlık Sorunları, Tıbbi Müdahaleler, Yüz ve Vücut Bakım	Sağlık	18

14	Arkadaş İlanları ve Kişisel İlanlar, Çevrimiçi Topluluklar, E-Posta ve Mesajlaşma, Romantik İlişkiler	Arkadaşlık - Sohbet	25
15	Başvuru Kaynakları, Bilim, Çeviri Araçları ve Kaynakları, Dil Kaynakları, Din ve İnanç, Geçmiş, İlaçlar, İslam, Kitaplar ve Edebiyat, Doğaüstü Olaylar, Sözlükler ve Ansiklopediler, Emeklilik, Vize ve Göçmenlik, Emeklilik, Sosyal Hizmetler	Bilgi Kaynağı	63
16	Dedikodu ve Magazin Haberleri, Dünya Haberleri, Gazeteler, Haberler, Siyaset, Televizyon ve Radyo Haberleri, Yerel Haberler	Gazete ve Haber	67
17	Çizgi Filmler, Film Referansları, Filmler, Radyo, Televizyon Kanalları, Televizyon Pembe Dizileri, Televizyon Suç ve Adalet Dizileri, TV Dizi	Tv - Sinema	86
18	Çevrimiçi Video, Fotoğraf ve Video Paylaşımı, Müzik Yayınları ve İndirilebilir Öğeler, Video Paylaşımı, Müzik ve Ses	Video ve Müzik Paylaşımı	77
19	Araç Ruhsatı ve Tescili, Araç Tekerlek ve Lastikleri, Ayakkabılar, Beslenme, Cep Telefonları, Eczacılık, Ev Aletleri, Ev Mobilyaları, Fiat, Ford, Hastaneler ve Tedavi Merkezleri, Hava Yolculuğu, İnşaat Malzemeleri ve Gereçleri, İSS'ler, Kablo ve Uydu Sağlayıcıları, Kuryeler ve Posta Hizmetleri, Makyaj ve Kozmetik Ürünleri, Nissan-Infiniti, Oteller ve Konaklama, Perakende Ticaret, Raylı Taşımacılık	İşletmeler	69
20	Astroloji, Yemek Tarifleri, Dans Müziği, El Sanatları, Fotografik Sanat, Güzellik ve Egzersiz, Hamilelik ve Annelik, Ebeveynlik, Kadınlara Özgü Alanlar, Kendi Kendine Yardım ve Motivasyon, Kulüpler ve Gece Hayatı, Nasıl Yapılır, Kendin Yap ve Uzman Yardımı, Sanat ve Eğlence,Seyahat, Fotoğrafçılığı, Şans Oyunları, Şarkı Sözleri ve Notaları, Şiir	Kişisel ilgi-beceri alanları	79
Toplam			1000

Bu veri dönüştürme işlemi ile çok sayıdaki kategorileri olan web siteleri, yönelik olduğu konu ve işlevleri dikkate alınarak benzerliklerine göre birleştirilmiştir. Bu nitelik, WEKA'da 20 farklı sınıf bazında nominal veri olarak değerlendirilecektir.

Erişim: Bu nitelik bir yüzde ifadesidir ve sürekli değişken veri özelliğindedir. Türkiye'deki farklı ziyaretçilerin değerinin (21 Milyon kişi) yüzde olarak kaçına erişildiğini göstermektedir. Örneğin, % 26'lık bir oranla ülkedeki, internet kullanıcılarının % 26'sına ulaşılabiliyor anlamına gelir.

Farklı ziyaretçiler (tahmini çerezler) ve Sayfa görüntüleme sayısı: Bu niteliklere ait verilerde işlem kolaylığı olması açısından sıfırlar (6 adet sıfır rakamı) atılmıştır. Değerler milyon seviyesindedir.

Sitede geçirilen ortalama süre: Bu çalışmada, sınıflandırma algoritmaları için seçilen 2 hedef nitelikten birisidir ve çalışmanın bundan sonraki bölümünde kısaca "ortalama süre" olarak geçecektir. Bu niteliğe ait veriler, dakika ve saniye cinsinden elde edilmiştir.

İki farklı birime ait olan bu değer WEKA'da işlem yapılabilmesi için tek birime (dakika) dönüştürülmüştür.

Ortalama ziyaret sayısı: Bu çalışmada WEKA'da sınıflandırma algoritmaları için seçilen diğer bir hedef niteliktir.

Sürekli değişken (sayısal) tipteki diğer 5 niteliğe ait basit istatistik Tablo 3 ile gösterilmektedir:

Tablo 3. Sürekli Değişken Niteliklerin Basit İstatistiği

Sürekli değişken tipteki nitelikler	Maksimum değer	Minimum değer	Ortalama değer	Standart sapma
Erişim (%)	74,8	0,4	1,70	3,64
Farklı ziyaretçiler (tahmini çerezler) (milyon)	60	0,24	1,32	2,90
Sayfa görüntüleme sayısı (milyon)	40000	0,42	59,80	1266,97
Ortalama süre (dk)	30	0,55	6,86	4,53
Ortalama ziyaret sayısı	58	3,3	6,32	3,49

Çalışmada kullanılan 5 farklı niteliğe ilişkin sürekli değişken verilerdeki değişimler, verilerdeki artışlar ve azalışlar incelenerek üç farklı sınıfta gruplanarak veri dönüştürme işlemleri tamamlanmıştır. Tablo 3 'den görüleceği gibi örneğin erişim niteliği %1'den az olanlar, %1 ve %3 arasında olanlar ve %3'ten fazla olanlar şeklinde 3 gruba ayrılmıştır.

Tablo 4. Sürekli Değişken Niteliklerin Veri Dönüşümü Sonrasında Gruplara Ayrılması

Sürekli değişken tipteki nitelikler	Veri dönüşümünden sonra	Veri sayısı	Toplam
Erişim (%)	[1-];[1-3];[3+]	567; 326; 107	1000
Farklı ziyaretçiler (tahmini çerezler) (milyon)	[1-];[1-3];[3+]	934; 53; 13	1000
Sayfa görüntüleme sayısı (milyon)	[50-];[50-100];[100+]	601; 368; 31	1000
Ortalama süre (dk)	[5-];[5-10];[10+]	413; 403; 184	1000
Ortalama ziyaret sayısı	[5-];[5-10];[10+]	431; 467; 102	1000

Verilerin WEKA'da Simgelenişi

Çalışmada verilerin algoritmalarla incelenmesi ve analiz edilmesi WEKA (Waikato Environment for Knowledge Analysis) yazılımının 3.6.4 versiyonu ile yapılmıştır. WEKA, makine öğrenmesi, veri madenciliği ve istatistiksel alanda kullanılan, açık kaynak kodlu Java tabanlı bir yazılımdır.

Veri dönüştürme işlemleri sonrasında analizde kullanılacak 1000 web sitesinin sahip olduğu 6 nitelik ve bu niteliklere bağlı veriler. Arff uzantılı dosya formlarına getirilerek WEKA platformuna yüklenmiştir.

Tablo 5. Doubleclick Ad Planner İle Elde Edilen Örnek Veriler

	Web Sitesi	Site kategorisi	Erişim (%)	Farklı ziyaretçiler (çerezler) (milyon)	Sayfa görüntüleme sayısı (milyon)	Sitede geçirilen ort.süre (dakika)	Ort. Ziyaret Sayısı
1	facebook.com	Sosyal Ağlar	74.8	60	40000	30	58
2	live.com	Arama Motorları	47.2	37	1100	7.5	24
3	msn.com	Web Portalları	23.8	20	250	5	14
4	mynet.com	Web Portalları	23.6	19	1000	11.16	18
5	meb.gov.tr	Devlet	23.6	19	1100	11.83	16

DoubleClick Ad Planer ile elde edilen verilerden ilk 5'i dönüşüm işlemleri öncesindeki formuyla Tablo 5'de verilmektedir. Veri dönüştürme işlemleri sonrasında, .arff dosya türüne dönüştürülmüş hali ile dosya içeriğinin ilk kısmından bir bölüm Şekil 2'de verilmektedir.

```
@relationveriseti

@attributeSite_turu {Sosyal_paylasim, Internet_portali, Resmi_kurum, Video_ve_muzik_paylasimi, Isletmeler, Ilanlar, Bilgi_kaynagi, Gazete_ve_haber, Oyun,e-Ticaret, Bilisim_teknoloji, Kisisel_ilgi_beceri_alanlari, Dosya_paylasimi, Forumlar, Ekonomi_bankacilik, Arkadaslik_sohbet, Tv_sinema, Spor, Saglik, Egitim}
@attributeErisim{3+,1-3,1-}
@attributeFarkliziy(cerez){3+,1-3,1-}
@attributeGoruntusayfasayi{100+,50-100,50-}
@attributeOrtSure{10+,5-10,5-}
@attributeOrtziyaret{10+,5-10,5-}

@data
Sosyal_paylasim,3+,3+,100+,10+,10+
Internet_portali,3+,3+,100+,5-10,10+
Internet_portali,3+,3+,100+,5-10,10+
Internet_portali,3+,3+,100+,10+,10+
Resmi_kurum,3+,3+,100+,10+,10+
```

Şekil 2. Verilerin dönüşümü yapıldıktan sonraki .arff dosya tipi şeklindeki formu

Bulgular

Makine Öğrenmesi Algoritmaları Performans Değerlendirme Ölçütleri

Makine öğrenmesi algoritmaları, bir uygulama üzerinde sınandığı zaman hangi oranda başarı elde edildiği bilinmesi istenir. Değerlendirme ve algoritmaların karşılaştırılması için birçok kavramdan yararlanılabilinir. Bu kavramlardan en çok kullanılanlar, doğruluk oranı, keskinlik, duyarlılık ve F-ölçütüdür. (Gencer ve ark. 2008).

Tablo 6. Karmaşıklık Matrisi Genel Formu

Karmaşıklık matrisi		Tahmini Sınıf	
		Pozitif	Negatif
Gerçek sınıf	Pozitif	TP	FN
	Negatif	FP	TN

Sınıflandırma işlemi sonucunda sınıflara göre tahmini yapılan veriler, gerçek sınıflarına göre yorumlandığında Tablo 6'da gösterilen 4 durumdan birisine ait olmaktadır. Makine öğrenme algoritmalarında performans değerlendirmesinde kullanılan kavramlar ve denklemler aşağıda sıralanmaktadır(7,8,9,10):

$$\text{Doğruluk} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FN} + \text{FP} + \text{TN}} \tag{7}$$

$$\text{Keskinlik} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{8}$$

$$\text{Duyarlılık} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{9}$$

$$\text{F Ölçütü} = \frac{2.\,\text{Duyarlılık}.\,\text{Kesinlik}}{\text{Duyarlılık} + \text{Kesinlik}} \tag{10}$$

Gözetimli Öğrenme Algoritmaları ile Analiz Sonuçları

Çapraz Doğrulama Testi Performans Sonuçları ve Karşılaştırması

1000 veri üzerinde iki farklı hedef nitelik için elde edilen çapraz doğrulama test sonuçları Tablo7'de verilmiştir.

Tablo 7.Gözetimli Öğrenme Algoritmaları Çapraz Doğrulama Testi Sonuçları

Çapraz Doğrulama - Cross Validation (n=10) [1000 veri]	Hedef Nitelik									
	Ortalama süre					Ortalama ziyaret sayısı				
	Keskinlik	Duyarlılık	F -Ölçütü	Doğru-Yanlış veri sayısı	Doğruluk (%)	Keskinlik	Duyarlılık	F- Ölçütü	Doğru-Yanlış veri sayısı	Doğruluk (%)
NaiveBayes	0.647	0.65	0.648	650-350	65	0.712	0.715	0.707	715-285	71.5
Bayes Ağı	0.646	0.649	0.647	649-351	64.9	0.718	0.72	0.712	720-280	72
Destek Vektör Makinesi	0.65	0.654	0.65	654-346	65.4	0.726	0.721	0.714	721-279	72.1
K En Yakın Komşu	0.643	0.642	0.642	642-358	64.2	0.694	0.697	0.694	697-303	69.7
ID3	0.652	0.65	0.65	637-343	63.7	0.705	0.704	0.704	692-291	69.2
C4.5	0.674	0.679	0.674	679-321	67.9	0.692	0.701	0.686	701-299	70.1

Gözetimli öğrenme algoritmalarında, n=10 için çapraz doğrulama yöntemi ile yapılan test sonuçlarında algoritma başarıları arasında büyük fark olmamakla birlikte hedef nitelik ortalama süre alındığında en yüksek doğruluk oranı % 67.9 ile C4.5 karar ağacı algoritmasında görülmüştür. C4.5 algoritmasında keskinlik, duyarlılık ve F-ölçütü değerlerinin hepsi de, diğer algoritmalara göre üstünlük sağlamıştır. Aynı hedef nitelik için ikinci en yüksek doğruluk oranı % 65.4 ile Destek Vektör Makinesinde elde edilmiştir. Her iki hedef nitelikte de en düşük performansı diğerlerine göre yaklaşık % 3'lük küçük bir başarı eksikliği ile ID3 algoritması göstermiştir.

Ortalama ziyaret sayısı hedef nitelik olması durumunda ise Destek Vektör Makinesi tüm performans kriterlerin de üstünlük sağlamıştır. % 72.1 doğruluk oranı sağlayan bu algoritma, keskinlik, duyarlılık ve F-ölçütünde sırasıyla % 72.6, %72.1 ve % 71.4 değerlerini sağlamıştır. Hemen ardından ikinci en yüksek performansı ise % 72 doğruluk oranı ise Bayes Ağı göstermiştir. Bayes Ağı diğer tüm kriterlerde de Destek Vektör Makinesi değerlerine yaklaşmıştır. Diğer gözetimli öğrenme algoritmalarında da bu başarı oranına yakın performanslar elde edilmiştir.

İncelenen hedef nitelikler açısından oluşan genel bir durum ise ortalama ziyaret sayısına ait algoritma performansının ortalama süre algoritma performansından daha yüksek çıkmasıdır. Ortalama % 5'lik bir fark, süreden ziyade ziyaret sayılarının bu analizlerde daha elverişli olduğunu göstermiştir. İki farklı hedef nitelik arasında en büyük fark Bayes ağında gözlenmiştir. Doğruluk oranında % 7.1 bir fark mevcuttur. En yakın değerlerin görüldüğü algoritma ise C4.5 algoritmasıdır. Doğruluk oranları arasındaki fark sadece % 2.2dir.

Eğitim ve Test Kümeleri Performans Karşılaştırması

Gözetimli öğrenme algoritmalarının yeni verilerdeki tahminleme başarısını gözlemlemek için ayrı ayrı yapılan 700 eğitim verisi ve 300 adet test verisi sonuçlarında

yer alan doğruluk oranları ve bu oranlar arasındaki farklar karşılaştırmalı olarak Tablo 8'de verilmektedir.

Tablo 8. Eğitim/Test Sonuçları Açısından Algoritmaların Karşılaştırılması

700 Eğitim / 300 Test Verisi	Hedef Nitelik					
	Ortalama süre			Ortalama ziyaret sayısı		
GözetimliÖğrenme Algoritmaları	Doğruluk oranı (%) (Eğitim kümesi)	Doğruluk oranı (%) (Test küme)	Bağıl Fark (%)	Doğruluk oranı (%) (Eğitim kümesi)	Doğruluk oranı (%) (Test küme)	Bağıl Fark (%)
NaiveBayes	68.71	64.66	4.05	72.57	70.33	2.24
Bayes Ağı	68.85	65	3.85	72.57	70	2.57
Destek Vektör Makinesi	70.71	65	5.71	74.71	71.33	3.38
K En Yakın Komşu	79.14	66	13.14	83.57	73.66	9.91
ID3	79.14	63.33	15.81	83.57	71	12.57
C4.5	73.42	68	5.42	71	69.66	1.34

Sonuçlara bakıldığında C4.5 ve Destek Vektör Makinesi algoritmaları test işleminde de öğrenme başarısını % 1.34 ila % 5.71 arasındaki hata payıyla koruduğu söylenebilmektedir. Diğer algoritmalardan K En Yakın Komşu ve ID3 algoritmaları eğitim verileri ile % 79.14 başarı sağlasa da test işleminde başarılarını % 9.91 ila % 15.81 aralığında kaybettiği gözlenmiştir.

Gözetimsiz Öğrenme Algoritmaları ile Analiz Sonuçları

Gözetimsiz öğrenme algoritmaları kullanılarak yapılan kümeleme analizinde, benzer özellikleri gösteren veriler aynı kümede toplanır ve verilere ilişkin bir fikir elde edilmeye çalışılır. Gözetimsiz öğrenme algoritmaları, gözetimli öğrenme algoritmalarının tersine modeli denetimsiz olarak öğrendikleri için niteliklerden herhangi birisi hedef nitelik değildir ve verilerin gözetimli öğrenmedeki gibi önceden belirlenen sınıflara atanması söz konusu değildir (Kırmızıgül, 2008). Böyle bir atama olmadığı içinde doğruluk vb. oranlar elde edilememektedir. Bu da verilerin doğru ayrım yapılıp yapılmadığının tespitini zorlaştırmaktadır. Ancak, kümelemede temel amaç küme içi benzerliklerinin maksimum ve kümeler arası benzerliklerin minimum olduğu küme sayısını bulmaktır. Bunun için farklı küme sayıları için oluşan kümelerdeki verileri inceleyerek karar vermek gerekir.

Eğitim ve Test Kümeleri Sonuçları ve Karşılaştırması

Bu çalışmada, web sitesi istatistikleri üzerinde uygulaması yapılacak gözetimsiz öğrenme algoritmalarının hem küme sayısını belirlemede, hem de algoritmaların kıyaslaması WEKA' daki test seçenekleri ile belirlenmiştir.1000 adet veriden 700'ü eğitim ve geri kalan 300'ü test için kullanılmıştır. Eğitim verileri ile WEKA' da" Usetraining set"

seçeneği, test için ise "Supplied test set" seçeneği kullanılmıştır. Girilen k küme sayısı ile iki veri seti için k adet kümenin veri sayılarına ait yüzde oranları bulunmuştur. Daha sonra bulunan bu k kümenin yüzde oranları eğitim-test olarak karşılaştırılmış ve bağıl farkları bulunduktan sonra ortalamaları alınmıştır. Bu işlem k=2 ve k=6 arasında her k sayısı için yapılmıştır ve ortalama sonuçlar Tablo 9' da verilmiştir.

Tablo 9.Gözetimsiz Öğrenme Algoritmalarının (700/300) Veri İçin Karşılaştırılması

700 Eğitim-300 Test verisi	k küme sayılarına göre oluşan kümeler arasındaki veri sayısının fark ortalaması (%)				
Gözetimsiz Öğrenme Algoritmaları	k=2	k=3	k=4	k=5	k=6
K-means	0	8.46	2.52	4.17	2.67
Hiyerarşik Kümeleme (En uzak komşu)	4.73	3.20	4.16	4.55	19.90

WEKA' da eğitim ve test verilerinde oluşan kümelerin veri sayıları ile tespit edilen yakınlık oranlarına gören az hata k=2 için % 0 hata ortalamasıyla olmuştur. k=2'den sonra en yakın oran k=4 küme için olmuştur. Buradaki hata ortalaması ise % 2.52 olmuştur. En fazla hata ortalaması ise % 8.46 ile k=3 için olmuştur.

WEKA' da Hiyerarşik kümelemelerden bağlantı tipi olarak Complete seçilerek En uzak komşu algoritması analiz edilmiştir. Hiyerarşik kümeleme algoritmaları ile eğitim ve test verileri ile oluşan kümeler arasındaki hata ortalamasının en az olduğu durum k=3 için % 3.20 ile olduğu gözlenmiştir. Bunun ardından en az hata ortalaması k=4 için gerçekleşmiştir.

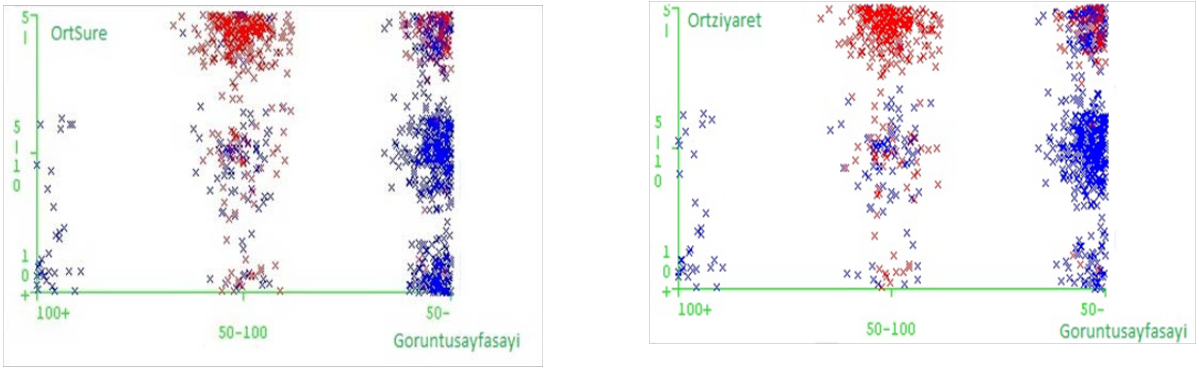
Gözetimsiz iki algoritma karşılaştırıldığında hata ortalamasının % 0 olduğu k=2 küme sayısı için test işleminde en isabetli kümeleme K-means algoritması ile gerçekleştiği görülmektedir. Bu bakımdan K-means algoritmasının, Hiyerarşik kümeleme algoritmasına göre üstünlük sağladığı söylenebilir. En uygun k sayısı tespit edildikten sonra çalışmada kullanılan veriler (1000 veri) WEKA' da k=2 kümeye ayrılmıştır.

K-means algoritması ile en ideal şekilde 2 kümeye ayrılan verilerin her nitelik için küme merkezini temsil eden sınıflar ve bu sınıfların kendi nitelikleri içindeki veri sayısına bağlı yüzde oranları Tablo 10'da verilmektedir.

Tablo 10. K-Means Algoritması Sonuçları: Küme Merkezleri Ve Veri Yüzdeleri

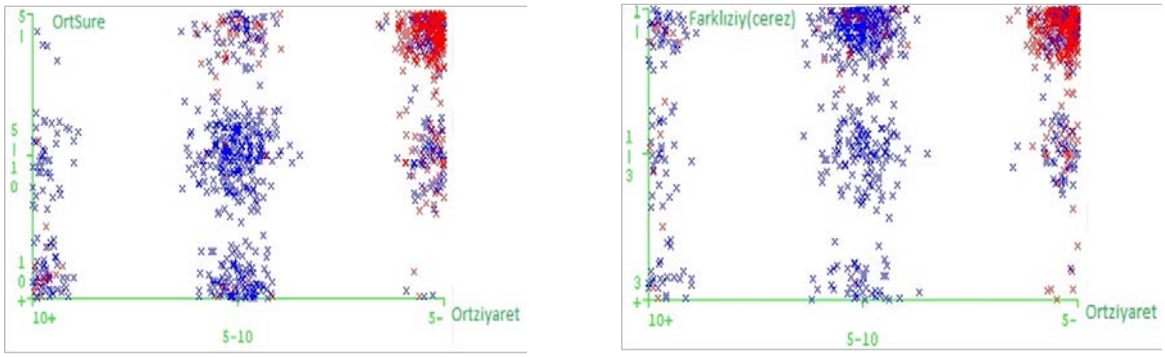
Nitelikler	Küme0		Küme1	
	Sınıf	Oran (%)	Sınıf	Oran (%)
Site türü (20 sınıf)	Oyun	10.51	Kişisel ilgi...	10.13
Erişim (3 sınıf)	1-3	38.09	1-	70.63
Farklı ziyaretçiler (3 sınıf)	1-	58.09	1-	89.52
Görüntülenen sayfa sayısı (3 sınıf)	50-	75.77	50-100	58.98
Ortalama süre (3 sınıf)	5-10	50.17	5-	67.99
Ortalama ziyaret sayısı (3 sınıf)	5-10	63.11	5-	73.34
Tüm verilerin kümelere dağılımı	618 (%62)		382 (%38)	
Withinclustersum of squarederrors: 2351.0				

K-means algoritması ile iki kümeye ayrılan 1000 adet verinin 618 tanesi Küme0 olarak mavi renkle, kalan 382 tanesi ise Küme1 olarak kırmızı renkle simgelenmiştir.



(a) (b)

Şekil 3.Web sitelerinin (a) Görüntülenen sayfa sayısı ve ortalama süre(b) Görüntülenen sayfa sayısı ve ortalama ziyaret sayısı niteliklerine göre dağılımları Şekil 3(a)'da web sitelerinin ortalama süre-görüntülenen sayfa sayısı niteliklerine göre dağılımlarına bakıldığında kümeleri ayıran nitelikler arasında ortalama sürenin diğer niteliğe göre az farkla baskın olduğu söylenebilmektedir. Küme1'in ortalama süre=5- hattında, Küme0'ın ise görüntülenen sayfa sayısı=50- hattında, ortalama süre=5-'ye kadar yoğunlaştığı görülmektedir. Şekil 3(b)'de de benzer bir grafik ortalama ziyaret sayısı-görüntülenen sayfa sayısı grafiğinde oluşmuştur. Verilerin ait oldukları kümeyi belirleyen etken, görüntülenen sayfa sayısından daha çok ortalama ziyaret sayısı niteliği olmuştur.



(a) (b)

Şekil 4. Web sitelerinin (a) Ortalama ziyaret sayısı ve ortalama süre (b) Ortalama ziyaret sayısı ve Farklı ziyaretçiler (tahmini çerezler) sayısı niteliklerine göre dağılımları

Şekil 4(a) 'da ortalama süre niteliği ve ortalama ziyaret sayısının oluşturduğu küme dağılımlarında kümelerin ayrımında ortalama ziyaret sayısının, ortalama süreye göre daha etkin olduğu söylenebilmektedir. Küme0 ortalama ziyaret sayısı için 5-10 sınıfında, ortalama sürenin 5-10 ve 10+ sınıfında yoğunlaşırken, Küme1 ise her iki niteliğinde 5-10 sınıfında yoğunlaşmıştır.

Şekil 4(b)'de ortalama ziyaret sayısının farklı ziyaretçiler (tahmini çerezler) niteliğine göre kümelerin ayrılmasında baskın olduğu görülmektedir. Bu ayrımın, daha çok ortalama ziyaret sayısının 5- sınıfından 5-10 sınıfına geçerken olduğu söylenebilmektedir. Verilerin dağılımda ise farklı ziyaretçiler (tahmini çerezler) niteliği 1 milyonun altında iken, verilerin önemli yoğunluğu koruduğu görülmektedir.

Sonuç

Bu çalışmada, Türkiye'deki web siteleri istatistikleri kullanılarak makine öğrenmesi algoritmaları üzerinde detaylı bir inceleme ve analiz yapılmıştır. Gözetimli öğrenme algoritmaları ile ortalama süre ve ortalama ziyaret sayısı parametreleri kritik hedef nitelik seçilmiş olup algoritmaların birbirleri ile performans karşılaştırmaları, sonuçlara ilişkin yorumlar yapılmıştır.

Gözetimli öğrenme algoritmaları ile yapılan sınıflandırma analizinde çapraz doğrulama yönteminde sonuçlar genellikle birbirine yakın çıkmıştır. Ortalama süre hedef nitelik seçildiğinde doğruluk oranı en fazla C4.5 algoritmasında görülmekte iken, ortalama ziyaret sayısında en fazla Destek Vektör Makinesinde görülmüştür. Böylece, bu alanda yapılacak olan çalışmalarda iyi bir başarı elde edilebileceği de gösterilmiştir. Diğer bir sonuç olarak da ortalama ziyaret sayısı hedef niteliği, her algoritmada ortalama hedef niteliğine göre daha yüksek performans göstermiştir. Bu da sınıflandırma modelleri için ortalama ziyaret sayısının önemini ortaya koymaktadır.

Ayrı bir test kümesi sınaması yapılan gözetimli öğrenme algoritmalarında ki ID3 ve K en yakın komşu algoritması, her iki hedef nitelikte de diğer algoritmalara göre farklı davranmıştır. Nitekim bu iki algoritmanın test kümesinde gösterdiği doğruluk oranları ile eğitim kümesi doğruluk oranları arasındaki fark daha fazladır. Bu da bu algoritmaların, sınıfı tahmin edilmesi istenen yeni bir veri için, iyi bir tahmin yapamayabileceğini göstermektedir. Web siteleri tıklama analizi için yanıltıcı sonuçlar ortaya koyabileceği sonucu ortaya çıktığından K en yakın komşu ve ID3 algoritmalarının bu tür analizlerde kullanılması dezavantajlı olabilecektir.

Gözetimsiz öğrenme algoritmalarında eğitim ve test veri kümeleri için K-means ve Hiyerarşik Kümeleme yönteminden En Uzak Komşu Algoritması kümeleme sonuçlarında eğitim ve test verileri için ayrı ayrı oluşan kümelerde yüzdelik oranlar karşılaştırılmıştır. En az fark oranı ortalaması k=2 küme için K-means algoritmasında gözlenmiş ve veriler 618-382 veri şeklinde dağılmıştır. 618 veri içeren küme niteliklerinde sınıflara bakıldığında, diğer kümeye oranla daha yüksek değerlere sahiptir. Farklı niteliklerin küme ayrımlarındaki etkisine grafikler yardımıyla bakıldığında ise ortalama ziyaret sayısı ve ortalama süre niteliklerinin diğer

niteliklere oranla daha belirleyici oldukları görülmüştür. Bu da web sitesi istatistiklerinde bu niteliklerin, verilerin hangi kümeye ait olmasında ağırlığı olduğunu ve öneminin büyük olduğunu göstermektedir. Bu iki nitelik arasında yapılan kümeleme kıyaslamasında ise, verilerin kümelere aitliğine daha çok etkiyi yapan ortalama süreye nazaran, ortalama ziyaret sayısıdır. Bu durum da sınıflandırma modellerindeki algoritmalara ait çapraz doğrulama sonuçlarını desteklemektedir.

Yapılan bu çalışma, web istatistiklerinin analizinde ve web ağ trafiğinin yorumlanmasında makine öğrenmesi algoritmalarının anlamlı parametre tespitinde kullanabilirliği ve yararlılığı araştırılmıştır. Verilerin değerlendirilmesi istenildiğinde sınıflandırma ve kümeleme modellerinde izlenecek yol ve yöntemleri inceleyerek yararlılığı yüksek parametreler tespit edilmiştir.

Kaynakça

Alpaydın, E.(2004). *Introductionto Machine Learning,* The MIT Press.

Amasyalı, m. F., bilgin, M., (2015), *Sekans Etiketleme Uygulamaları için Makine Öğrenmesi Yöntemlerinin Karşılaştırılması Comparison of Machine Learning Methods for the Sequence Labelling Applications*.

Anderberg, M. R. (1973), *Cluster Analysis for Applications, New York: AcademicPress.*

Bülbül,Ş., Güler, M.F., Kandemir, A.Ş., (2009), *Propensity Skor Uygulamalarında Kümeleme Analizinin Test Amaçlı Kullanımı*,172.

Coşkun, C., Baykal, A., *Veri Madenciliğinde Sınıflandırma Algoritmalarının Bir Örnek Üzerinde Karşılaştırılması*, Dicle Üniversitesi, Fen Fakültesi Matematik Bölümü, Diyarbakır.

Cristianni, N.,Shawe-Taylor, J.(2000). *An IntroductiontoSupportVectorMachinesand OtherKernel-Based Learning Methods*, UK: Cambridge UniversityPress.

Çetın, N. M., Haciomeroglu, M. (2013). *Survey of GPU Accelerated Data Clustering Algorithms*. AJIT-e, 4(11), 19.

Demirci, D.A. (2007). *Destek Vektör Makineleri İle Karakter Tanıma*, Yüksek Lisans Tezi, Yıldız Teknik Üniversitesi.

Dünder, E., Cengiz, M. A., Koç, H., Savaş, N. (2013). *Bayesci ağlarda risk analizi: Bankacılık sektörü üzerine bir uygulama*. Erzincan üniversitesi sosyal Bilimler Enstitüsü Dergisi, 6(1), 1-14.

Gencer, C., Akbulut, S., Kızılkaya Aydoğan, E. (2008), *Churn Analysis AndCustomer SegmentationOf A CosmeticsBrand Using Data MiningTechniques*, Journal of Engineeringand Natural Sciences, Sigma. Vol./Cilt 26 Issue/Sayı 1.

Günes, A., Yiğit, T. (2012, April). *Handwritten digit recognition with accelerated Support Vector Machines*. In Signal Processing and Communications Applications Conference (SIU), 2012 20th (pp. 1-4). IEEE.

Habibi, Y., Sheisi, G. H., Abdi, H. (2015). *Voltage Instability Detection in Power System Using Support Vector Machine (SVM).*

Johnson, R. A., Dean W. W. (1999). *AppliedMultivariate Statistical Analysis(Fourth Editon).*, New Jersey: PrenticeHall, UpperSaddleRiver.

Kavzoğlu, t., Çölkesen, İ. (2010). *Karar Ağaçları İle Uydu Görüntülerinin Sınıflandırılması.* Harita Teknolojileri Elektronik Dergisi, 2(1), 36-45.

Kaya, Ç., & Yıldız, O. (2014). *Makine öğrenmesi teknikleriyle saldırı tespiti: Karşılaştırmalı analiz.*

Kırmızıgül Çalışkan, S.,Soğukpınar, İ. (2008), *Knn: K-Means Ve K En Yakın Komşu Yöntemleri İle Ağlarda Nüfuz Tespiti*, TmmobEmo 2.Ağ Ve Bilgi Güvenliği Ulusal Sempozyumu.

Lee, C. H. (2015). *A gradient approach for value weighted classification learning in naive Bayes. Knowledge-Based Systems*.

Monz, C.,*Machine Learning for Data Mining, Week 6: Clustering*.

Quinlan J.R. (1993). *C4.5: Programs for Machine Learning*, San Mateo, CA: Morgan Kaufmann.

Özkan, Y. (2008). *Veri Madenciliği Yöntemleri* (1.Baskı). İstanbul: Papatya Yayıncılık.

Öztemel, E.(2006). *Yapay Sinir Ağları* (2.Baskı). İstanbul: Papatya Yayıncılık

Saini, I., Singh, D., Khosla, A. (2013). *QRS detection using K-Nearest Neighbor algorithm (KNN) and evaluation on standard ECG databases*. Journal of advanced research, 4(4), 331-344.

Silahtaroğlu, G. (2008).*Veri Madenciliği Kavram ve Algoritmaları*. İstanbul: Papatya Yayıncılık.

Sorias, S. (2014). *Psikiyatrik Tanıda Betimsel ve Kategorik Yaklaşımların Kısıtlılıklarını Aşmak: Bayes Ağlarına Dayalı Bir Öneri*.

SPSS.(1999).*AnwerTreeAlgorithmSummary*. SPSS White Paper, USA.

Uzun, E. (2007). *İnternet tabanlı bilgi erişimi destekli bir otomatik öğrenme sistemi*.

Yıldırım, S. (2003). *Tümevarım Öğrenme Tekniklerinin C4.5'in İncelenmesi*, Yüksek Lisans Tezi. İstanbul: İTÜ.

http://tr.wikipedia.org/wiki/Makine_%C3%B6%C4%9Frenimi, E.T: 07.05.2014

http://www.google.com/support/adplanner,https://adwords.google.com/da/DisplayPlanner/Home ET:10.04.2011

http://www.turkeyforum.com/satforum/archive/index.php/t635294.html,E.T:21.05.2014

http://www.damlakaraman.com.tr/google-doubleclick-nedir, E.T:18.06.2014