Hakemli - Süreli - Bilimsel Dergi
İNTERNET UYGULAMALARI ve YÖNETİMİ
Journal of
INTERNET APPLICATIONS and MANAGEMENT
2015 / 6(2)

Bir E-Ticaret Sitesi Kullanıcı Hesaplarında Şifre Yapılarının Birliktelik Kuralları ile İncelenmesi

The analysis of passwords structures in an e-commerce site user accounts by using association rules

Onur DOĞAN [1], onur.dogan@deu.edu.tr

Geliş Tarihi/Received: *23.11.2015****; Kabul Tarihi/Accepted:*** *28.12.2015* ***doi:*** *10.5505/iuyd.2015.29491*

Bireysel hesapların güvenliği için hesap şifrelerinin gücü büyük önem taşır. Bir şifrenin gücü; uzunluğu, karmaşıklığı ve kolay tahmin edilebilir olmaması ile ölçülebilir. Veri madenciliği; veri setlerinden, veri sahibi için kullanışlı ve anlaşılır bilgilerin çıkarılması süreci olarak tanımlanabilir. Veri madenciliği teknikleri sayesinde, klasik metotlarla elde edilemeyen bilgiler bilgisayar yazılımları vasıtasıyla elde edilebilir. Organizasyon ve bireylerin verinin toplanması ve depolanması için yaptıkları eylemler ve geliştirdikleri teknolojiler, veriler kullanışlı hale getirilmediği müddetçe bir anlam ifade etmezler. Veri kümelerinden bilgi çıkarma amacıyla kullanılan veri madenciliği model ve teknikleri her geçen gün geliştirilmeye devam etmektedir.Bu çalışmada, bir e-ticaret sitesi kullanıcılarının şifre yapılarına ilişkin bazı istatistiksel çıkarımlar ortaya konmuştur. Buna göre, kullanıcı şifre uzunlukları, 4 karakter ile 12 karakter arasında değişmektedir. Şifre uzunluk ortalaması 7,1 olup, şifrelerin %53'ünün yalnızca bir karakter kullanılarak oluşturulduğu saptanmıştır. Buradan yola çıkarak şifrelerin büyük bir çoğunluğunun yeterli güvenliğe sahip olmadığı tespit edilmiştir. Ayrıca, kullanıcıların şifre ve bazı diğer bilgileri ile oluşturulan veri kümesi, veri madenciliği tekniklerinden birliktelik analizlerine tabi tutularak değişkenler arası kullanışlı kural tabanlı bilgiler elde edilmeye çalışılmıştır. Elde edilen kuralların site yöneticileri için kullanışlı olabileceği ifade edilebilir.

Anahtar Kelimeler: *Şifre güvenliği, Şifre gücü, Veri madenciliği, Birliktelik analizleri*

Jel Kodları: *C89, L86.*

The strength of a password is very important for the security of an individual account. The strength of a password can be measured by its length, complexity and predictability. Data mining can be defined as a process of knowledge (useful and understandable for the owner of the data) extraction from the data. Owing to data mining tools and techiques, some knowledge -which is not obtainable by using conventional methods- can be extracted by using computer softwares. Developing technologies and doing some other activities for data collection and data storage by organizations and individuals will not make sense unless the data becomes useful. Data mining models and techniques which were used for information extraction from data sets continue to improve every day. This study has revealed some statistical inferences on an e-cosmmerce site users' individual passwords and structures of passwords. Accordingly, the users' password lengths changes between 4 - 12 characters. The average password length was found 7,1 and also 53% of passwords generated using only one character. It could be said that many users' passwords are not safe enough. In addition, a data set which is generated by users' passwords and other information about users, has been subjected to association analysis to obtain some useful rule-based knowledge. It could be told these rules are useful for the decision makers.

Keywords: *Password security, Password strength, Data mining, Association analysis*

Jel Codes: *C89, L86.*

[1] Öğr. Gör. Dr., Dokuz Eylül Üniversitesi, İzmir MYO

1. GİRİŞ

Günümüzde hala çoğu sistem için şifre yazarak kimlik doğrulama işlemi yapılmaktadır. Kişiler sosyal medya hesaplarına, e-posta adreslerine, alışveriş sitelerinde ve diğer birçok farklı amaçla kullandıkları sitelerindeki hesaplarına bireysel bir şifre yazarak ulaşmaktadır. Kişilerin şifre tercihlerinin, bireysel veri güvenlikleri açısından taşıdığı önem açıktır.

US-CERT (United States Computer Emergency Readiness Team), şifre seçimi ve güvenliği konusunda 2013 yılında yayınladığı raporda, şifre seçimi tavsiyelerini aşağıdaki gibi sıralanmıştır:

- Kolay ulaşılabilir ya da tahmin edilebilir kişisel bilgilerinizi şifre olarak kullanmayınız.
- Herhangi bir dilin sözlüğünde bulunan bir kelimeyi şifre olarak kullanmayınız.
- Karmaşık şifreleri hatırlamak için bir hatırlatıcı geliştiriniz.
- Hem büyük hem küçük karakter kullanınız.
- Şifrenizi hem özel karakter hem de sayılardan oluşturunuz.
- Mümkünse şifre hatırlatıcı kullanınız.
- Farklı sistemlerde farklı şifreler kullanınız.

Bir şifrenin güç değerinin; şifrenin uzunluğu, karmaşıklığı ve tahmin edilememesi kriterlerinin bir fonksiyonu olarak tanımlanabileceği söylenebilir (US-CERT, 2013). Sistem yöneticileri kullanıcıların veri ve bilgi güvenliklerini sağlamak için kullanıcı hesaplarında şifre belirleme konusunda bazı kriterler koymaktadırlar. Örneğin, bazı sitelerde belli bir karakter sayısının altında şifreler kabul edilmemekte veya kullanıcılardan hem büyük hem küçük harf hem de sayı içeren şifreler oluşturmaları istenmektedir.

Bireylerin ve organizasyonların daha doğru kararlar alabilmesinde sahip oldukları verileri anlamlı ve kullanışlı bilgiye dönüştürmeleri kritik bir süreçtir. Özellikle günümüzde elektronik ortamda farklı biçimlerde ve büyük çapta veri kümeleri üretilirken, bu veri kümeleri içerisinden kullanışlı bilgi elde edilmesi verimliliğin ve etkinliğin arttırılması açısından büyük önem kazanmıştır. Son yıllarda akademik alanlarda ve iş dünyasında, birçok proje, makale, araştırma vb. çalışmada kullanılan veri madenciliği model ve metotları; veri kümeleri içerisinden anlamlı ve kullanışlı bilgi elde edilmesi amacına hizmet etmektedirler.

Bu çalışmada bir e-ticaret sitesindeki kullanıcıların şifre ve bazı diğer özelliklerinden oluşan bir veri seti temel istatistiksel analizlere tabi tutularak kullanıcı şifrelerinin genel yapısı ortaya konmuştur. Ayrıca veri setinden, veri madenciliği tekniklerinden biri olan birliktelik algoritmaları yardımı ile kullanıcıların şifrelerine ve özelliklerine dair kullanışlı kural tabanlı bilgiler çıkarılmaya çalışılmıştır.

2. LİTERATÜR TARAMASI

Kullanıcı şifrelerinin güvenlik değerleri ile ilgili ilk çalışmalardan biri Morris ve Thompson (1979) tarafından yapılmıştır. Bu çalışmada 3289 şifre üzerinde araştırma yapılmış ve şifrelerin %86'sı çok zayıf bulunmuştur. Zayıf şifreler ya çok kısa ya da büyük ve küçük harf

kombinasyonlarından yalnızca birini içermektedir. Riddle vd. (1989) yaptıkları çalışmada katılımcıların %88'inin şifrelerinin 4 karakter veya daha az karakterden oluştuğunu gözlemlemiştir. Ayrıca, bu kısa şifrelerin %44'ünün de aynı zamanda bir İngilizce kelimeye denk geldiği ve tahmin etmesi kolay olduğu ortaya konmuştur. Brown vd. (2004)'nin 218 öğrencinin şifreleri üzerinde yaptığı araştırmaya göre; katılımcıların şifre karakter uzunlukları ortalaması 4,45 olarak bulunmuştur. Riley (2006)'in 315 kişilik örneklem üzerinde yaptığı çalışmaya göre ise katılımcıların %85,7'sinin şifrelerinin yalnızca küçük harften oluştuğu belirlenmiştir. Bir şifrenin kriptografik gücü, uzunluğu (N) ile şifrenin belirleneceği alfabedeki karakter sayısı (C) değerlerinden N log2 C değerinin elde edilmesi ile oluşturulan bir ölçüm ile belirlenebilir. Burr vd., (2006)'nin çalışmaları da bu güç ölçümüne farklı yaklaşımları içeren bir çalışmadır. Florencio ve Herley (2007); bireysel şifre kullanma alışkanlıklarına yönelik çalışmalarında da geçmiş çalışmalara benzer bir şekilde kullanıcıların düşük kalitede şifreler kullandıkları ve şifrelerin birden fazla hesapta kullanılması ve şifre unutulması olaylarının sıkça yaşandığı belirtmiştir. Haque, Wright ve Scielzo (2013) çalışmalarında kişisel şifreleri; finansal sitelerde kullanılanlar, içerik-haber sitelerinde kullanılanlar, kişiye özgü sitelerde (sosyal medya hesapları gibi) kullanılanlar ve rastgele sitelerde kullanılanlar olarak ayırmış ve kişilerin şifrelerini kırmaya çalışmışlardır. Buna göre; kişilerin finansal sitelerde kullandıkları şifrelerde daha özenli oldukları (en az bir büyük harf ya da sayı içeren şifre belirledikleri) ortaya çıkmıştır. Ayrıca, kendi şifre kırma metotlarının şifre içerisindeki bir kelimede anlamsal benzerliğin çıkış noktası olduğunu belirtmişlerdir. Liu, Hong, Pi (2014)'nin Çinli internet kullanıcıları üzerinde yaptıkları çalışmada, kullanıcıların 4 farklı sitedeki şifre bilgilerini toplamışlardır. Buna göre kullanıcıların ortalama şifre uzunluklarının 4 site için 7,74 ile 9,45 arasında değiştiği, ayrıca 4 sitede de kullanıcılardan şifrelerini yalnızca sayıdan oluşturanların diğer şifre kombinasyonları ile şifre belirleyenlerden daha fazla oldukları tespit edilmiştir.

Veri madenciliği; genellikle büyük veri setlerinin, veri sahibi için yararlı ve anlaşılır olacak biçimde, umulmadık ilişkiler yakalamak ve özgün bir biçimde özetlemek için analiz edilmesidir (Hand vd., 2001: 2). Cabena vd. (1998)'ne göre veri madenciliği; büyük veri tabanlarından bilgi çıkarımı için kullanılan ve makine öğrenimi, örüntü tanıma, istatistik, veri tabanları, görselleştirme gibi alanlardan teknikleri bir araya getirendisiplinler arası bir alandır. Han ve Kamber (2001)de veri madenciliğinin birçok disiplini kapsadığına dikkat çekmekte ve önceki tanımdaki alanlara ek olarak; yüksek performanslı hesaplama, yapay zekâ, bilgi tabanlı sistemler, yapay sinir ağları, bilgi çıkarımı gibi alanlardan tekniklerin de veri madenciliği kapsamında kullanıldığını belirtmektedir. Han Kamber ve Pei (2012); "verilerden bilgi madenciliği (knowledge mining from data)" kavramının veri madenciliğinde yapılan işlere daha uygun olsa da uzun olduğu için, "bilgi madenciliği (knowledge mining)" kavramının ise büyük veri kavramına atıf yapılmadığı için tercih edilmediğini vurgulamışlardır. Ayrıca, "bilgi çıkarımı", "veri/desen analizi", "veri arkeolojisi" ve "veri tarama" gibi ifadelerin de veri madenciliği ile benzer anlamlarda olduklarını belirtmişlerdir.

Bunlara ek olarak, veri madenciliği ile veri tabanlarından bilgi keşfi kavramları her ne kadar bazen birbirleri yerine kullanılan kavramlar olsalar da aralarında fark olduğu belirtilmelidir. Dunham (2002); veri tabanlarından bilgi keşfini verilerden yararlı bilgi ve desen bulma süreci olarak tanımlarken, veri madenciliğini ise veri tabanlarından bilgi keşfi sürecinde

ortaya çıkan, bilgi ve desen keşfi için algoritma kullanımı olarak tanımlamaktadır. Han Kamber ve Pei (2012) ise bu iki kavram arasındaki farkı tanımlarken; veri madenciliğinin, veri tabanlarından bilgi keşfi sürecinin ana basamağı olduğuna dikkat çekmektedirler.

Veri tabanlarından bilgi keşfi aşağıdaki adımları aşağıdaki gibidir (Han&Kamber, 2000: 6);

1. Veri Temizleme
2. Veri Birleştirme
3. Veri Seçimi
4. Veri Dönüşümü
5. Veri Madenciliği
6. Veri Değerlendirme
7. Bilgi Sunumu

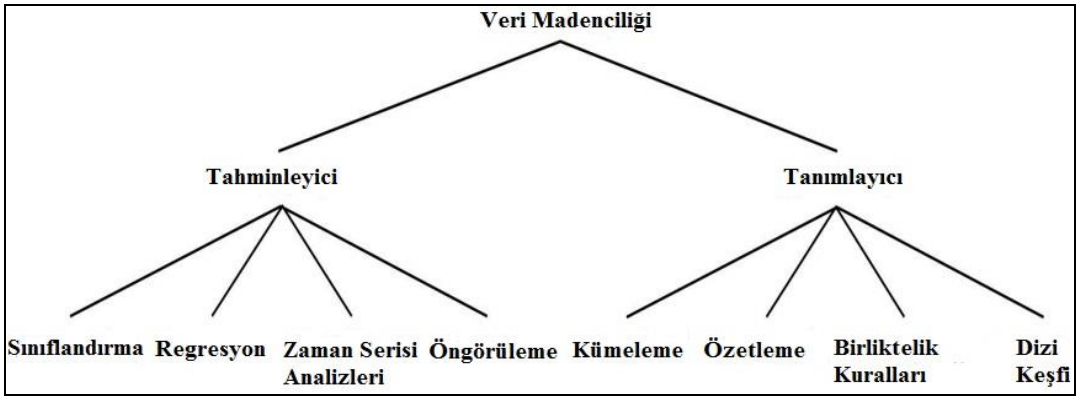
Bu basamaklar tanımlanırken, veri madenciliğinin; veri desenleri ortaya çıkarmak için uygulanan akıllı metotların uygulandığı ana süreç olduğuna dikkat çekilmiştir.

Literatürde veri madenciliğine ait birçok tanımlama mevcut olmakla beraber, basit bir anlatım ile büyük hacimli veri kümelerinden anlamlı, anlaşılır ve kullanışlı bilgilere ulaşmak için kullanılan tüm araç, teknik ve yöntemlerin veri madenciliği alanına girdiğini söylemek yanlış olmayacaktır.

Veri madenciliği tekniklerinin işlevleri(yetenekleri) aşağıdaki gibi sıralanabilir (Larose, 2005:11);

- Tanımlama
- Kestirim
- Öngörü
- Sınıflama
- Kümeleme
- Birliktelik

Dunham (2002) ise veri madenciliği sınıflandırmasında öncelikle tahminleyici ve tanımlayıcı olarak ki başlık altında yapmıştır. Tahminleyici işlevler; sınıflandırma, regresyon, zaman serisi analizleri ve öngörüleme olarak sıralanmıştır. Tanımlayıcı işlevler ise kümeleme, özetleme, birliktelik kuralları ve dizi keşfi (*sequence discovery*) olarak sıralanmıştır.



Şekil 1. Veri Madenciliği Modelleri ve İşlevleri (Dunham, 2002)

Veri kümesi içerisinden birliktelik kuralları çıkarılması, genel bir ifadeyle veriyi oluşturan kayıtlara ait değişkenlerden iki yönlü olarak (bir hedef değişken seçilmeden) anlamlı kurallar çıkarılması olarak tanımlanabilir. Bu çalışmadaki verilerden, karar vericilerin ya da yöneticilerin öngörmediği anlamlı kurallar elde edilmeye çalışılacağı için veri madenciliği metotlarından birliktelik kuralları çıkarımı kullanılacaktır.

3. METODOLOJİ

Birliktelik kuralları ilk defa 1993 yılında Agwaral, Imielinski ve Swami tarafından ortaya konulmuştur. Araştırmacılar bu tip kuralların ortaya çıkarılması için bir takım düzenlemeler ve bir algoritma önerisi yapmışlardır. Birliktelik kurallarının genel bir formu; $X \rightarrow Y$ (X ve Y veri kümesindeki öğeler olmak üzere) şeklinde verilebilir.

Birliktelik kuralları, sınıflandırma kurallarına benzemektedir. Sınıflandırma algoritmaları, veri setini analiz ederek hedef bir değişkenin hangi diğer değişken değerlerine göre oluşmuş olduğunu ortaya koyarak, kurallar tanımlar ve yeni gelecek bir verinin bu hedef değişkenin hangi değerini alacağının bulunmasına yardım eder. Söz gelimi bir banka verdiği kredilerin risk değerlendirmesini yapıyor olsun. Bankanın sahip olduğu veri kümesinde önceki müşterilerinin "yaş", "gelir durumu", "meslek", "eğitim durumu" ve "kredinin geri ödenip ödenmediği bilgisi" var ise; kredinin geri ödeme bilgisi hedef değişken olarak tayin edilmelidir. Bu takdirde; karar verici belli bir güven değeri ile yüksek yaş grubunda, gelir durumu düşük olan kişiye kredi verilmemesi gibi bir kural ortaya koyabilir. Birliktelik kurallarında ise böyle bir hedef değere ihtiyaç yoktur.

Witten ve Frank (2005), birliktelik kuralları ile sınıflandırma arasında, birliktelik kurallarının, herhangi bir sınıf yerine; tüm değişkenler arası kombinasyonlar ile kurallar oluşturması haricinde hiçbir fark olmadığını belirtmektedirler.

Birliktelik algoritmaları, büyük veri kümelerinden değişkenler arasında her yönde (belirli bir hedef değer ortaya konulmadan) ve her tipte ilişkinin ortaya çıkarılması için kullanılan algoritmalardır.

İlk birliktelik algoritması Agwaral, Imielinski ve Swami tarafından isimlerinin baş harfi ile AIS adı altında ortaya konulmuştur. Bunu, Houtsma ve Swami (1993) SETM (Set-Oriented Mining) algoritması takip etmiştir. 1994 yılında Agwaral ve Srikant bugün sıklıkla kullanılan Apriori algoritmasını, bunun yanı sıra AprioriTid ve bu iki algoritmanın birleştirilmesi ile oluşturulmuş AprioriHybrid algoritmasını ortaya koymuşlardır (Agwaral ve Sikant, 1994). Ayrıca; Scheffer (2001) PredectiveApriori birliktelik kuralı algoritmasını geliştirmiştir.

Bu çalışmada kullanılacak olan veri madenciliği yazılımı olan WEKA, Waikato Üniversitesi'nde geliştirilmiş, açık kaynak kodlu ve ücretsiz bir veri madenciliği yazılımıdır (Hall, vd. 2009). Programı internet sitesinden indirmek mümkündür. Ayrıca, programda kullanılan algoritmalar ile ilgili bilgiler yazılım içerisinde verilmektedir. WEKA'da birliktelik analizi için kullanılan algoritmalar; *Apriori*, *Filtered Associator*, *FPGrowth*, *Generalized Sequantial Patterns*, *Predictive Apriori* ve *Tertius* olarak sıralanabilir.

Kullanılan algoritmalar sonrası bulunan kuralların anlamlılığı ya da kullanışlılığının ölçülmesi gerekmektedir. Veri madenciliği sonrası bulunan birliktelik kurallarının işe yararlılığının ölçümü, ilginçlik ölçümleri (measures of interestingness) adı verilen teknikler ile yapılmaktadır. Literatürde birliktelik kurallarının ilginçlik değerlerini bulmak için birçok

teknik geliştirilmiştir. Bu çalışmada kullanılması nedeni ilginç değerlerinden güven (confidence), destek (support) ve doğruluk (accuracy) değerleri kullanılmıştır.[2]

Birliktelik kuralları için birer ilginçlik ölçüsü olan güven ve destek değerleri aşağıdaki biçimde hesaplanır (Agwaral, Imielinski ve Swami, 1993: 207-216);

D: Veri tabanı, t_i: Veri tabanındaki kayıtlar, $D = \{t_1, t_2, \ldots, t_m\}$

X, Y: Kurallardaki öğeler (Öncül veya Ardıl)

$X \rightarrow Y$, X'in öncül, Y'nin ardıl olduğu bir kural $(X \cap Y = 0)$ olmak üzere;

Destek:

$$supp(X) = \frac{|\{t \in D; X \subset t\}|}{|D|} \quad (1)$$

Güven(Güç):

$$\mathrm{conf}(X \rightarrow Y) = \frac{\mathrm{supp}(X \rightarrow Y)}{\mathrm{supp}(X)} \quad (2)$$

Destek ve güven değerlerinin ne ifade ettiklerini anlamak adına basit bir örnek vermek gerekirse; bir internet sitesinden alışveriş yapan 100 kişinin hangi ürünleri aldıkları bilgisine sahip olduğu varsayılsın. D; bu 100 kişinin kaydı olarak tanımlanır bu takdirde, $|D|$=100 olur. X; bu siteden cep telefonu alma işlemini gösteriyor ve 40 kişi cep telefonu almış ise; $|X|$=40 olur. Y işlemi cep telefonu kılıfı alma işlemini tanımlar ve cep telefonu alan 40 kişinin 30 tanesi aynı zamanda kılıf alıyor ise; bu durumda, "Eğer bir kişi cep telefonu alıyor ise kılıf da alacaktır $(X \rightarrow Y)$" ifadesi bir birliktelik kuralı olarak ortaya konabilir. Bu kuralın destek değeri; supp(X)=40/100=0,4 ve güven değeri; conf$(X \rightarrow Y)$=30/40=0,75 olarak bulunur.

Güven ve destek değerlerinin birlikte kullanılması ile elde edilen kestirim doğruluğu değeri (predictive accuracy) ilk defa Scheffer (2001) tarafından ortaya konmuştur. Bu değer ilginçlik değeri veya kestirim doğruluk değeri adlarıyla adlandırılabilir ve PredictiveApriori algoritması başta olmak üzere bazı algoritmalarda kuralın güvenirlik düzeyini gösteren bir değer olarak kullanılır.

4. UYGULAMA

Bu çalışmada bir e-ticaret sitesinin kullanıcılarının bireysel hesap şifreleri üzerinden analiz yapılacaktır. Site; konfeksiyon ürünleri başta olmak üzere, diğer tekstil ürünleri, kozmetik, aksesuar vb. geniş bir yelpazede hizmet veren, kullanıcıların kendi hesapları ile internet üzerinden alışveriş yapabildiği bir sitedir. Kullanıcılar siteye; doğum tarihi, cinsiyet, adres gibi bilgileri doldurarak, kendi belirleyecekleri kullanıcı adları veya mail adresleri ile ve oluşturdukları bir şifre ile üye olmaktadırlar.

Çalışmanın veri kümesinde 10050 müşteriye ait isim, soy isim, bulundukları şehir, kullanıcı adı, şifre, yaş, cinsiyet bilgileri bulunmaktadır. Bu çalışmanın amacı bu veri tabanındaki

[2] Diğer ilginçlik değeri ölçümleri için Tan, Kumar ve Srivastava (2004), Hahsler ve Hornik (2007), Geng ve Hamilton (2006) çalışmaları incelenebilir.

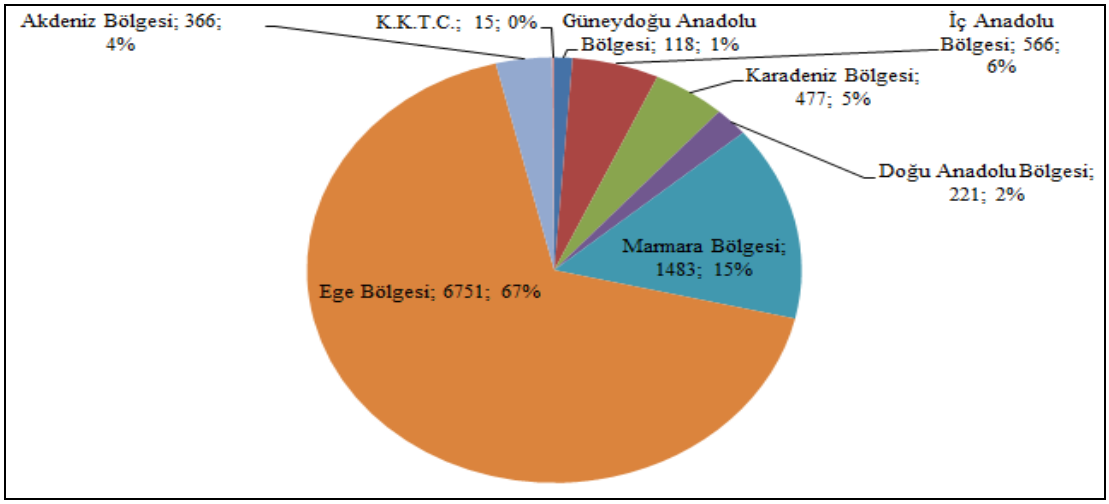
internet alış verişi yapan kişilerin cinsiyet, şifre yapıları değişkenlere ilişkin genel profillerini ortaya koymak ve veri kümesinde anlamlı ve kullanılabilir kurallar çıkarmaktır.

4.1. Çalışmada Kullanılan Veri Setinin Yapısı ve Analiz İçin Hazırlanması

Ham veri kümesinde bulunan 10050 veri, hem kayıt bazında hem de değişken bazında incelenerek bir takım işlemlerden sonra veri madenciliği için hazır hale getirilmiştir. Öncelikle, eksik ve hatalı bilgiler içeren 53 kayıt analiz dışı bırakılmıştır. Tüm değişkenleri belirli olan 9997 kişilik veri kümesi analiz için uygundur. Birliktelik kuralları kategorik veriler ile çalışmaktadır. Bu nedenle, sürekli değer alabilen değişkenler birliktelik kuralları için kategorileştirilmelidir (Bramer, 2007: 187). Veri setindeki yaş değişkeni 18 ile 72 arasında değişmektedir. Yaş değişkenine ilişkin kategorileştirme işlemi TÜİK (2013) yaş grupları baz alınarak yapılmıştır. Bu sınıflandırma; "0-14","15-17","18-24","25-44","45-64","65-84" ve "85 ve üstü" olarak tanımlanmışsa da, çalışmadaki veri setinde 18 yaş altı ve 85 yaş üstü kişi bulunmadığından yedi kategori yerine, yaş değişkeni dört kategori altında toplanmıştır. Bireylerin adres bilgileri Türkiye'deki 7 coğrafi bölge ve bunlara ek olarak Kuzey Kıbrıs Türk Cumhuriyeti (K.K.T.C) olmak üzere 8 farklı gruba ayrılmıştır. Kullanıcıların şifreleri[3], 4 karakter ile 12 karakter arasında değişmektedir. Kategorileştirme amacıyla şifre uzunluğu değişkeni, 4-6 arası grup "kısa", 7-9 arası grup "orta", 10-12 arası grup "uzun" olarak adlandırılarak uzunluk adı altında, şifre uzunluğunu dilsel bir değişken olarak ifade eden yeni bir grup oluşturulmuştur. Bunlara ek olarak, veri tabanındaki şifreleri; harfler, sayılar ve özel karakterler oluşturmaktadır. Şifreler üzerinden karmaşıklık adı altında yeni bir grup tanımlanmıştır. Bir şifre harf, sayı ve özel karakter değerlerinden yalnızca birini içeriyorsa 1, herhangi ikisini içeriyorsa 2 ve her üç türü de içeriyorsa 3 değerini almaktadır.

4.2. Verilere İlişkin Frekans Analizleri

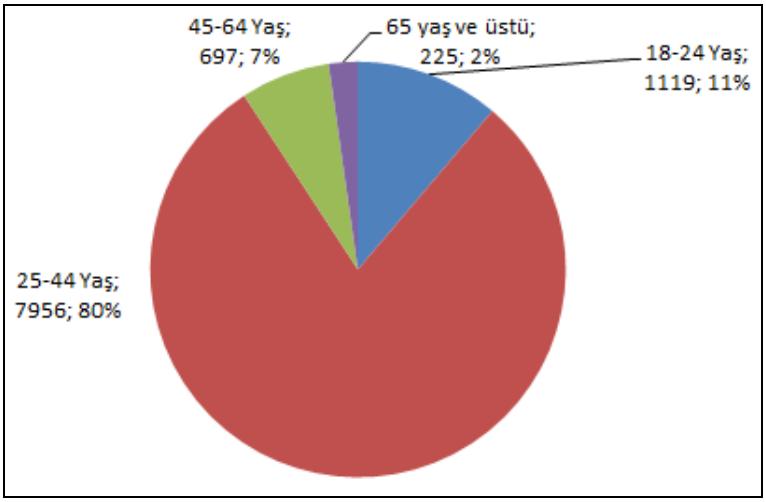
Bu çalışmadaki veri kümesindeki kişilerin 3346 tanesi (%33) kadın, 6651 tanesi (%67) erkektir. Kişilerin bulundukları bölge bilgileri Şekil 3'te gösterildiği gibidir.



Şekil 2. Kullanıcıların Bulundukları Bölge Bilgileri

[3] Firma, kişisel güvenlik nedeni ile şifreleri olduğu gibi değil, bir değişikliğe tabi tutarak tarafımla paylaşmıştır. Buna göre şifre uzunluğu aynı kalmış, şifredeki bir harf bir başka harfle, bir sayı bir başka sayıyla ve aynı biçimde bir özel karakter bir başka özel karakter ile değiştirilerek verilmiştir.

Katılımcıların 366'sı Akdeniz Bölgesinden, 15'i K.K.T.C. 'den, 118 tanesi Güneydoğu Anadolu Bölgesinden, 566 tanesi İç Anadolu Bölgesinden, 477 tanesi Karadeniz Bölgesinden, 221 tanesi Doğu Anadolu Bölgesinden, 1483 tanesi Marmara Bölgesinden ve en büyük çoğunluğu oluşturan 6751 tanesi Ege Bölgesindendir.



Şekil 3. Kullanıcıların Yaş Bilgileri

Veri setindeki değişken grupları incelendiğinde, yaş bazında e-ticaret sitesine kayıtlı kişilerden 44 yaş alt kitle veri kümesinin büyük bir çoğunluğunu oluşturmaktadır (7956 kişi). İnternet üzerinden alışveriş yapan kişilerin çoğunlukla genç bireyler olduğu düşünüldüğünde, bu istatistik anlamlıdır. 18-24 yaş grubunda 1119 kişi, 45-65 yaş grubunda 697 kişi ve 65 yaş üstü grupta 225 kişi bulunmaktadır.

Şifre bilgilerine ilişkin istatistikler Tablo 1 'de sunulmuştur.

Tablo 1. Şifre Bilgilerine İlişkin İstatistikler

Şifre Karmaşıklığı	1	5320
	2	3909
	3	768
Şifre Uzunluğu	Kısa	3711
	Orta	4901
	Uzun	1385
Kullanıcı Adı	Mail Adresi	9956
	Diğer	41

Bu siteye üye olan bireylerin şifre uzunluğunun önemli bir kısmının (%53) şifre karmaşıklık değerinin 1 olduğu görülmektedir. Şifrelerini harf, sayı ya da özel karakterlerden yalnızca bir grubu kullanarak oluşturan 5320 kullanıcıdan; 3115 tanesi yalnızca harf ile 2198 tanesi yalnızca sayı ile ve 7 tanesi yalnızca özel karakter ile şifrelerini oluşturmuşlardır. Şifrede farklı grupların kullanımının hesap güvenliğini arttırdığı bir gerçektir. Kullanıcıların büyük

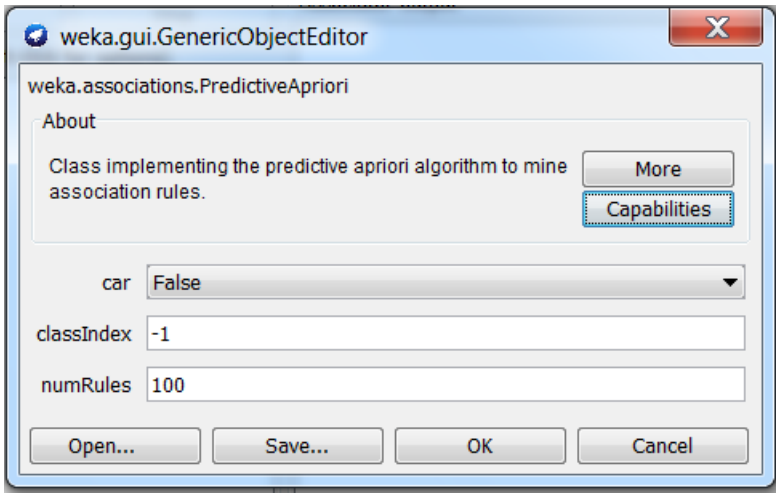
çoğunluğunun bunu önemsemediği belirtilebilir. Bir şifrenin güvenliğinin bir diğer kıstası ise şifre uzunluğudur. Kullanıcıların %37'si kısa denebilecek (4-6 karakter) bir şifre kombinasyonuna sahipken, orta uzunlukta (7-9 karakter) şifreye sahip olanlar %49 ve uzun (10-12 karakter) uzunlukta şifreye sahip olanlar %14 oranındadır. Frekans analizinden ortaya çıkan bir başka ilginç bilgi, kullanıcıların hemen hemen hepsinin kullanıcı adı olarak mail adresini kullanmalarıdır. Kullanıcıların yalnızca 41 tanesi e-posta adresi yerine kendilerine özgü bir kullanıcı adı belirlemişlerdir. Kişilerin belirlediği bir kullanıcı adı, kişilerin kullanıcı adı olarak e-posta adreslerini kullanmalarından tahmin edilebilirlik ya da ulaşılabilirlik açısından daha güvenlidir. Kullanıcılarının şifrelerinin ortalama uzunluğu 7,1 olarak tespit edilmiştir. Ayrıca, kullanıcıların hem harf, hem sayı hem de özel karakter kullanarak şifrelerini oluşturabilecekleri düşünülürse, şifre oluştururken bu üç karakteri de kullananların oranı tüm kullanıcıların yalnızca %8'idir. Kullanıcıların yarısından fazlası yalnızca bir karakter tipini kullanarak şifre oluşturmaktadır. Kullanıcı bilgilerinden elde edilebilecek bir başka önemli bulgu ise kullanıcıların neredeyse tamamına yakınının kullanıcı adı olarak mail adresini kullandıklarıdır.

4.3. Birliktelik Analizi Bulguları ve Çıkarımları

Kullanıcı adı değişkeninde mail adresi kullanımının baskınlığı nedeni ile analiz dışında bırakılmıştır. Veri seti bu düzenlemelerden sonra Weka programının birliktelik kural çıkarımı algoritmalarından PredictiveApriori algoritmasına tabi tutulmuştur.

Predictive Apriori birliktelik kuralı algoritmasında; destek ve güven değerleri birlikte ele alınarak kestirim doğruluğu (*predictiveaccuracy*) adı altında tek bir ölçü oluşturulur. Bu kestirim doğruluğu ölçüsü birliktelik kurallarının üretilmesinde kullanılır. Weka yazılımında bu algoritma, kullanıcıya en iyi "n" adet birliktelik kuralını verir(Aher&Lobo, 2012: 50).

Weka'da PredictiveApriori algoritması için bazı parametrelerin seçildiği işlem penceresi Şekil 4'te gösterilmiştir.



Şekil 4. Weka'da Predicti ve Apriori Algoritması Editörü

"car" değeri true veya false değerlerini alabilmektedir. Eğer bu değer true değerini alırsa genel birliktelik kuralları yerine sınıflandırma birliktelik kuralları kullanılacaktır demektir. Weka'da PredictiveApriori algoritması kural çıkarım sayısının (numRules) ön tanımlı değeri

100'dür. classIndex kısmı sınıf değişkeninin indeksi olarak tanımlanmıştır. Bu değerin, -1 olarak bırakılması durumunda son değişken sınıf değişkeni olarak tanımlanacaktır demektir.

Analizde ön tanımlı değer 100 olarak bırakılmış algoritmadan en iyi 100 kuralı çıkarması istenmiştir. Ayrıca herhangi bir değişken sınıf olarak atanmamış, genel birliktelik kuralı çıkarılması hedeflenmiştir.

Bütün veri madenciliği model ve metodolojilerinin her tipinde, her aşamada insan etkileşimine ihtiyaç vardır. Sürekli izleme, geçerlilik ve diğer değerlendirme ölçümleri insan analistler tarafından yapılmalıdır (Larose 2005: 10-11, 20).

Elde edilen kurallar gözden geçirilmiş, bazı kurallar düşük doğruluk (accuracy) değerleri nedeniyle[4] bazı kurallar çok az sayıda madde özelliğini kapsadığı için, bazı kurallar ise yüksek güvenirlik değerlerine sahip olsalar da herhangi bir anlam ifade etmedikleri için kural listesinden çıkarılmıştır. Bu eleme sonrası veri setinden aşağıdaki kurallar elde edilmiştir.

```
bolge=GAB karmasiklik=2 uzunluk=kisa 19 ==> yas=2 19   acc:(0.98923)

cinsiyet=erkek bolge=Icanadolu yas=3 uzunluk=kisa 12 ==> karmasiklik=1 12   acc:(0.97442)

cinsiyet=kadin karmasiklik=3 uzunluk=uzun 25 ==> bolge=Ege 25   acc:(0.90101)

cinsiyet=erkek bolge=DAB karmasiklik=2 yas=1 5 ==> uzunluk=orta 5   acc:(0.90101)

cinsiyet=kadin bolge=Marmara yas=1 45 ==> uzunluk=kisa 45   acc:(0.90101)

cinsiyet=erkek bolge=Karadeniz uzunluk=orta 160 ==> yas=2 140   acc:(0.85843)

cinsiyet=kadin karmasiklik=3 uzunluk=kisa 78 ==> bolge=Ege 68   acc:(0.84313)

bolge=GAB karmasiklik=2 yas=1 13 ==> uzunluk=orta 13   acc:(0.83571)

karmasiklik=4 yas=2 12 ==> bolge=Ege 11   acc:(0.82335)
```

Şekil 5.İnceleme Sonrası Birliktelik Kuralları

Birliktelik analizi sonucu elde edilen kurallarda yapılan eleme işleminden sonra dokuz adet anlamlı ve kullanışlı kural bulunmuştur. Kuralların ne anlama geldiğinin anlaşılması için bazıları incelenecek olursa;

1. kuralda; Güneydoğu Anadolu Bölgesinde ikamet eden, şifre karmaşıklık değeri 2, şifre uzunluk değeri kısa olan kişiler %98 doğruluk değeri ile 25-44 yaş grubundadırlar.

2. kural incelenecek olursa; cinsiyeti erkek olan, İç Anadolu'da oturan, 45-64 yaş arası erkek kullanıcılardan şifre uzunluğu kısa olan kullanıcıların, aynı zamanda şifre karmaşıklık değerleri de 1'dir (yalnızca bir karakter grubu ile şifre oluşturanlar). Kuralın öncül kısmındaki niteliklere uyan 12 kişinin, 12'si de kuralın ardıl kısmında aynı niteliğe (şifre karmaşıklığı 1 değeri) sahiptirler.

[4] 0,80 altındaki doğruluk değerleri çıkarılmıştır. Sınır değer olarak bu değerin alınması araştırmacının tercihidir.

Diğer kurallar da benzer şekilde yorumlanabilir. Kurallar e-ticaret sitesi için yeni hesap oluşturacak bireyler için bir takım yönlendirici bilgiler verilmesi konusunda yardımcı olabilir.

5. SONUÇ & TARTIŞMA

Çalışmaya konu olan e-ticaret sitesi kullanıcılarının ortalama şifre uzunluğu 7,1 olarak tespit edilmiştir. Bu ortalama değerin Riddle vd. (1989) ve Brown vd. (2004) çalışmalarındaki ortalama uzunluğu üzerinde olduğu gözlemlenmiş, ancak Liu, Hong, Pi (2014)'nin çalışmalarındaki 4 farklı şirket kullanıcılarının ortalama şifre uzunluklarının altındadır. Kullanıcıların şifre uzunluklarını arttırıcı zorlamalar site yöneticileri tarafından yapılabilir. Morris ve Thompson (1979) çalışmaları ile Riley (2006) çalışmasında kullanıcıların önemli bir bölümünün yalnızca harf kombinasyonuyla, Liu, Hong, Pi (2014) çalışmalarında ise kullanıcıların büyük çoğunluğunun yalnızca sayı kombinasyonu ile şifre oluşturduklarının tespit etmişlerdir. Bu çalışmada şifrelerinde bir karakter tipi kullananlardan yalnızca harf ile şifre oluşturanların sayısı, yalnızca sayı ile oluşturanların sayısından fazladır. Ancak buradaki asıl önemli nokta kullanıcıların yarısından fazlasının (%53) yalnızca bir karakter tipini tercih etmiş olmalarıdır.

Kullanıcıların önemli bir kısmının kullanıcı adı olarak herhangi bir isim belirlemeyip, e-postalarını kullanıcı adı olarak kullandıkları tespit edilmiştir. Kötü amaçlı kullanım amacıyla e-posta adres bilgilerine ulaşmanın, kişinin belirleyeceği bir kullanıcı adına ulaşmaktan daha kolay olduğu açıktır. Kullanıcı adı olarak bir başka isim belirlenmemesi, kısa şifreler ve yalnızca bir karakter tipi kullanılarak oluşturulan şifreler güvenlik zafiyeti oluşturabilir. Sitenin kullanıcı hesaplarının güvenliğini sağlaması için şifre uzunluğu, karmaşıklığı ve kullanıcı isimleri konusunda güvenliği arttırıcı kısıtlayıcı önlemler alması gerekmektedir yorumu yapılabilir.

Çalışmada etik kurallar gereği kullanıcı şifreleri değiştirilerek alınmıştır. Bu nedenle kullanıcı şifrelerinde anlamlı kelimelerin ya da kişi için önemli sayısal değerlerin (doğum günü tarihi, yaş vb.) kullanılıp kullanılmadığı tespit edilememiştir. Şifre gücünün tahmin edilebilirlik kriteri çalışmanın verileri için test edilememiştir. Bu durum çalışmanın kısıtları arasında belirtilmelidir.

Çalışmada elde edilen kurallardan biri; cinsiyeti kadın olan, Marmara bölgesinde ikamet eden, 18-24 yaş grubundaki kişilerin %90 doğruluk değeri ile şifrelerinin kısa olacağıdır. Site şifre belirleme aşamasından önce kuralın sol tarafını sağlayan bilgilere sahip bir kullanıcı için ekrana, uzun bir şifre belirlemeleri hususunda uyarı ibaresi çıkarabilir. Böylece bu özelliklere sahip kullanıcıların kısa şifre oluşturmaları engellenebilir.

Çalışmada kişilerin şifrelerinin güvenlik değerlerinin genel anlamada düşük olduğu ortaya konmuş, ayrıca sitenin hesap güvenlik ile ilgili biriminin sahip olunan veriler üzerinde uygulanabilecek bazı veri madenciliği teknikleri ile kullanışlı bilgiler çıkarabilecekleri gösterilmiştir.

KAYNAKÇA

Agwaral R., Imielinski, T. & Swami, A., (1993). Mining Association Rules Between Sets of Items in Large Databases. *Proceedings of the 1993 ACM SIGMOD Conference on Management of Data*, 207-216, Washington DC, USA.

Agwaral R. & Srikant, R. (1994). Fast Algorithms for Mining Association Rules. *Proceedings of the 20th VLDB Conference Santiago, Chile*, http://rakesh.agrawal-family.com/papers/vldb94apriori.pdf (Erişim Tarihi: 01.07.2015)

Aher, S. B. & Lobo, L. M. R. J. (2012). A Comparative Study of Association Rule Algorithms for Course Recommender System in E-Learning. *International Journal of Computer Applications*, 39(1), 48-52.

Brown, A.S., Bracken, E., Zoccoli, S. & Douglas, K. (2004). Generating and Remembering Passwords. *Applied Cognitive Psychology*, 18, 641-651.

Bramer, M. (2007). *Principles of Data Mining*. Springer-Verlag London.

Burr, W. E., Dodson, D. F. & Polk. W. T. (2006). *Electronic Authentication Guideline*. In NIST Special Publication 800-63.

Cabena P., Hadjinian P., Stadler R., Verhees J., & Zanasi A. (1998). *Discovering Data Mining: From Concept to Implementation*. PrenticeHall, Upper Saddle River, NJ.

Dunham, M. H. (2002). *Companion Slides for The Text by Dr. M. H. Dunham, Data Mining, Introductory and Advanced Topics*. Prentice Hall. https://deepalipawar.files.wordpress.com/2015/09/dataminingintroductiondifferent.pdf (Erişim Tarihi: 25.06.2015)

Florencio, D. & Herley, C. (2007). A Large Scale Study of Web Password Habits. *WWW 2007/Track: Security, Privacy, Reliability, and Ethics*, Session: Passwords and Phishing, 657-665.

Geng, L. & Hamilton, H. J. (2006). Interestingness Measures for Data Mining: A Survey. *ACM Computing Surveys*, 38(3), Article 9, 1-32.

Hall M., Frank E., Holmes, G., Pfahringer, B., Reutemann, P. & Witten I. H. (2009). The WEKA Data Mining Software: An Update. *SIGKDD Explorations*, 11(1), 10-18.

Han, J. & Kamber, M. (2001). *Data Mining, Concepts and Techniques*. Morgan Kaufmann Publishers.

Han, J., Kamber, M. & Pei, J. (2012). *Data Mining, Concepts and Techniques.* Morgan Kaufmann Publishers.

Hand, D., Mannila, H. & Smyth, P. (2001). *Principles of Data Mining*. MIT Press, Cambridge, MA.

Haque, S. M. T., Wright, M. & Scielzo, S. (2013). A Study of User Password Strategy for Multiple Accounts. *CODASPY 2013 - Proceedings of the 3rd ACM Conference on Data and Application Security and Privacy*, 173-175.

Houtsma, M. & Swami, A. (1993). *Set-Oriented Mining of Association Rules*. Research Report RJ 9567, IBM Almaden Research Center, San Jose, California.

Hahsler, M. & Hornik, K. (2007). New Probabilistic Interest Measures for Association Rules. *Intelligent Data Analysis*, 11(5), 437-455.

Larose, D. T. (2005). *Discovering Knowledge in Data - An Introductionto Data Mining.* John Wiley & Sons Inc., New Jersey, USA.

Liu, Z., Hong, Y. & Pi, D. (2014). A Large-Scalestudy of Web Password Habits of Chinese Network Users. *Journal of Software*, 9(2), 293-297.

Morris, R. & Thompson, K. (1979). Password Security: A Case History. *Communications of the ACM*, 22(11), 594-597.

Riddle, B. L., Miron, M. S. & Semo, J. A. (1989). Passwords in Use in A University Time Sharing Environment. *Computer Security*, 8(7), 569-578.

Riley, S. (2006). Password Security: What Users Know and What They Actually Do. http://psychology.wichita.edu/surl/usabilitynews/81/pdf/Usability%20News%2081%20-%20Riley.pdf (Erişim Tarihi: 01.08.2015)

Scheffer, T. (2001). Finding Association Rules That Trade Support Optimally Against Confidence. *In Proc. of the 5th European Conf. On Principles and Practice of Knowledge Discovery in Databases*, 424-435.

Tan, Pang-Ning, Kumar, V. & Srivastava, J. (2004). Selecting The Right Objective Measure for Association Analysis. *Information Systems, 29*, 293-313

TÜİK, (2013). Adrese Dayalı Nüfus Kayıt Sistemi (ADNKS), http://www.tuik.gov.tr/PreTablo.do?alt_id=1059 (Erişim Tarihi: 19.07.2015)

US-CERT (2013). Choosing and Protecting Passwords, https://www.us-cert.gov/ncas/tips/ST04-002 (Erişim tarihi: 12.07.2015)

WEKA, (2015). http://www.cs.waikato.ac.nz/ml/weka/ (Erişim Tarihi: 03.12.2015).