Hakemli - Süreli - Bilimsel Dergi
İNTERNET UYGULAMALARI ve YÖNETİMİ
Journal of
INTERNET APPLICATIONS and MANAGEMENT
2014 / 5(2)

İstenmeyen Elektronik Posta (Spam) Tespitinde Karar Ağacı Algoritmalarının Performans Kıyaslaması

The performance benchmark of decision tree algorithms for spam e-mail detection

Eyüp AKÇETİN [1], e.akcetin@gmail.com

Ufuk ÇELİK [2], ucelik001@gmail.com

Geliş Tarihi/Received: *15.01.2015****; Kabul Tarihi/Accepted:*** *11.03.2015* ***doi:*** *10.5505/iuyd.2014.43531*

Bu çalışmanın amacı istenmeyen elektronik postaların (spam) tespiti için veri madenciliği yöntemlerinden karar ağaçları algoritmalarının performanslarının kıyaslanarak doğruluk ve sınıflandırma modeli oluşum zamanı açısından en uygun olanının tespit edilmesidir. İstenmeyen elektronik postaların sınıflandırılması için gerekli olan veriler Kaliforniya Üniversitesi makine öğrenmesi veri setlerinden alınan 4601 adet elektronik posta ile sağlanmıştır. Veri madenciliği yöntemlerinden 12 farklı karar ağacı WEKA makine öğrenmesi yazılımı kullanılarak, 10 katlı çapraz doğrulama ile veri setinde istenmeyen elektronik postalar (spam) sınıflandırılmıştır. Bu sınıflandırmanın performansı, alıcı işlem karakteristiği analizi yapılarak belirlenmiştir. Bu çalışmada, istenmeyen elektronik postaların (spam) tespiti için karar ağaçlarının performansı incelendiğinde, 12 sınıflandırıcının doğruluk oranlarının %94.68 ile %91 arasında değiştiği tespit edilmiştir. Yapılan çalışmada, performans sonuçlarına göre rastgele orman algoritmasının %94.68 doğruluk oranı ile en iyi sınıflandırma başarısını elde ettiği tespit edilmiştir. Bu algoritmanın 4601 elektronik posta için sınıflandırma modeli oluşturma zamanı 2.11 saniye olup, yoğun bir elektronik posta alışverişi sisteminde istenmeyen elektronik postaları (spam) hızlı bir şekilde ayırt edebileceği anlaşılmıştır.

Anahtar Kelimeler: *Elektronik posta, spam, karar ağaçları, veri madenciliği, makine öğrenmesi.*

Jel Kodları: *C80, D81, Y10.*

The objective of this study is to determine the most convenient decision tree method in terms of accuracy and classification built time by comparing the performance of decision tree algorithms with the purpose of identifying the spam e-mails. The data were gathered from one of the datasets of University of California machine learning datasets including 4601 e-mails for the classification of spam. The spam e-mails were classified utilizing 10 fold cross validation by using WEKA machine learning software involving 12 different decision trees. The performance of this classification was found by implementing the principle component analysis. It was found that the performance of decision trees on determining spam e-mails showed accuracy rate ranging between 91% and 94.68%. Random Forest algorithm was found to be the best classifier with the accuracy rate of 94.68%. It was understood that this algorithm can classify spam e-mails quickly in a hectic e-mail exchange system because the classification built time of the algorithm is 2.11 seconds for the 4601 e-mails.

Keywords: *E-mail, spam, decision trees, data mining, machine learning.*

Jel Codes: *C80, D81, Y10.*

[1] Yard. Doç. Dr., Balıkesir Üniversitesi, Bandırma Denizcilik Fakültesi (Yazışılan yazar)
[2] Öğr. Gör., Balıkesir Üniversitesi, Gönen MYO

1. GİRİŞ

Hızla yayılan İnternet kullanımının beraberinde getirdiği sorunlardan en önemlisi bilgi suistimali ve İnternet dolandırıcılığıdır. Öte yandan tanıtım amaçlı elektronik postalar, tüketicileri ve/veya müşterileri rahatsız edici boyutlara ulaşabilmektedir. Bu ve benzeri durumlar, bilgisayar bilimcileri, istenmeyen elektronik postaların (spam) filtrelenmesini sağlayan algoritmalar geliştirmeye zorlamıştır. Yapılan literatür taramasında bu amaçla, istenmeyen elektronik postaların *(spam)* sınıflandırılması için veri madenciliği yöntemlerinin (Fawcett, 2003) başarılı bir şekilde kullanıldığı görülmektedir. Veri madenciliği, bir veri kümesinden alınan bilgileri analiz ederek anlamlı sonuçlar çıkarmak ya da tahmin etmek için kullanılan tekniklerdir. İstenmeyen elektronik posta tespiti için veri madenciliği yöntemlerinden yapay sinir ağları (Wu & Tsai, 2009), karınca kolonisi optimizasyonu (El-Alfy, 2009) veya kural tabanlı (Ruano-Ordás vd., 2013) algoritmalar kullanılarak spam filtreleme yapılmıştır. Spam filtreleme için karar ağaçları da kullanılmaktadır (Gaikwad & Halkarnikar; L. Shi vd., 2012).

Karar ağaçları, giriş verilerinin bir sınıflandırma veya kümeleme algoritmasıyla, tüm elemanların aynı sınıf etiketine ait olana kadar ağaç dallanmasına benzer şekilde alt gruplara ayrıştırma işlemidir (Orhan, 2012). Ayrıştırma için entropi, bilgi kazancı veya doğruluk oranı gibi ölçümler kullanılır. Karar ağaçları sınıflandırma, özellik seçimi, özellik çıkarımı ve karar kurallarının oluşturulması için kullanılabilir. Ayrıca farklı algoritmalar ile melez (hibrit) teknikler kullanılarak daha iyi sonuçlar alınabilir. Karar ağaçları başarılı bir performans ortaya koyan, düşük maliyetli, güvenilir, uygulaması kolay, anlaşılabilir ve yorumlanabilir olması sebebiyle çok tercih edilen veri madenciliği metodudur. Bu nedenle veri madenciliği yöntemlerinden karar ağaçları en sık kullanılan algoritmalardandır.

1.1. Literatür Taraması

İnternetin gelişmesi ile birlikte internet üzerinden dolandırıcılık ve bilgi suistimalleri de artacaktır. Bu artışa bağlı olarak bu tür problemlerle mücadele yöntemleri de gelişip dönüşecektir. Dolayısıyla gelecekte istenmeyen postaların *(spam)* filtrelenmesi önemli bir araştırma konusu olacaktır (Guzella & Caminhas, 2009).

Elektronik posta her ne kadar güçlü bir iletişim aracı olsa da kötü niyetli kullanımlara açık bir sistemdir. İstenmeyen elektronik postalar sadece elektronik posta kullanıcıları için değil aynı zamanda dijital güvenlik ve dünya ekonomisi için büyük bir sorun teşkil etmektedir. Küresel firmaların kampaya faaliyetleri olumsuz etkileye bilmekte hatta bazen piyasa değerlerini düşürebilmektedir. Bu nedenle bu tür elektronik postaların tespit edilip engellenmesi büyük önem arz etmektedir (Laorden vd., 2012).

Spam tespiti sadece elektronik postalarda değil aynı zamanda cep telefonlarına gönderilen reklam mesajlarının içinde yer almaktadır. Chan vd. 2015 yılında yaptığı çalışmada, reklam amaçlı gönderilen toplu mesajları ve Chicago'nun 20 en iyi oteli hakkında yorum içeren mesajları incelemiştir. Yapılan bu çalışmada reklam amaçlı gönderilen 5574 mesajdan 4827 adeti yasal ve doğru mesaj iken 747 adetinin spam olduğu tespit edilmiştir. Öte yandan oteller hakkında müşteri yorumlarını içeren 1600 mesajın 800'ü doğru iken 800'nün spam olduğu tespit edilmiştir (Chan vd., 2015).

Günden güne çoğalan istenmeyen elektronik postalar *(spam)* internet kullanıcılarına büyük sorunlar oluştururken ağ kaynaklarının verimsiz kullanılmasına yol açmaktadır. İstenmeyen elektronik postaların tespitinde makine öğrenmesi ve istatistiksel filtreleme yöntemleri kullanılsa da bu tür postaların tespitinde başarılı bir metot henüz yoktur. Dolayısı ile yapılan her çalışma kendine özgü yöntem ve metotları kullanmakta ve/veya test etmektedir (Ying vd., 2010).

Sosyal medyanın birçok konu üzerinde son derece belirleyici olduğu düşünüldüğünde, işletmelerin istenmeyen sahte elektronik postalar *(spam)* gibi sosyal medyada yer alan sahte profilleri de tespit edip onlara karşı önlem alması gerekmektedir. Sahte profil sahte mesaj, sahte elektronik posta, sahte resim ve haberler üreterek işletmelere ciddi ekonomik zararlar verebilmektedir (Ahmed & Abulaish, 2013).

1.2. Amaç

İşletmelerin en önemli araçlarından biri de elektronik postalardır. Bu haberleşme kanalının güvenliği ve verimliliği için istenmeyen elektronik postaların *(spam)* ekonomik ve kısa süre içerisinde tespit edilip bertaraf edilmesi gerekmektedir. Bilgi toplumlarında, bilgilerin manipülasyonu sonucu işletmeler büyük ekonomik kayıplarla yüzleşebilmektedir. Bu nedenle bu tür durumlarla etkin bir şekilde mücadele etmek, gerekli önlemleri etkin şekilde almak günümüzde zorunluluk haline gelmiştir. Bu noktada cevaplanması gereken en önemli soru; hangi yöntem en hızlı ve en ekonomiktir? Bu çalışmada, makine öğrenmesi türlerinden karar ağaçlarının ROC analizi yöntemi ile en verimlisi, algoritma model oluşum zamanına göre en hızlısı tespit edilmeye çalışılmıştır.

Bu bağlamda, istenmeyen elektronik postaların *(spam)* tespitinde en uygun algoritmanın tespit edilmesi için veri madenciliği yöntemlerinden 12 farklı karar ağacı algoritmasının performansları; doğruluk, hassasiyet ve kesinlik oranları sınıflandırma modelinin oluşum zamanına göre kıyaslanmıştır. İstenmeyen elektronik postaların tespiti içinde veri madenciliği yöntemleri kullanılırken, elektronik postaların niteliği veya hangi kuruluştan elde edildiği algoritma performansını etkileyen bir faktör değildir. Bu sebeple kullanılan elektronik postalardan elde edilen kelimelerin veya kullanılan işaretlerin frekanslarına göre sınıflandırma yapılmıştır.

Yapılan literatür taramasında istenmeyen elektronik postaların *(spam)* ve diğer bilgi suistimallerinin tespitinde dünyada kullanılan standart bir metodun olmadığı tespit edilmiştir. Bu çalışmanın amacı istenmeyen elektronik postaların *(spam)* tespitinde kullanılan yöntemlerden hangisinin en iyi olduğunu belli parametreler üzerinden tespit etmektir.

1.3. Materyal

Günümüzde işletmeler, elektronik postalarını, çeşitli stratejik nedenlerden dolayı, paylaşmaktan kaçınmaktadırlar. Bu nedenle bu çalışma kapsamında yapılacak analizlerde, Kaliforniya Üniversitesinin veri madenciliği için oluşturduğu açık kaynak kodlu veri tabanları kullanılmıştır.

İstenmeyen elektronik posta tespitinde karar ağaçlarını test etmek için, Kaliforniya Üniversitesi makine öğrenmesi veri setlerinden (Bache & Lichman, 2013) Spambase isimli, 1999 yılında Hewlett-Packard laboratuarlarından elde edilen 4601 adet elektronik posta

verileri kullanılmıştır. Bu postalarda 57 adet nitelikten ilk 48 tanesi postalardaki kelimelerin frekanslarını, 49'dan 54'e kadar 6 nitelik postalarda kullanılan ";", "(", "[", "!", "\$" ve "\#" gibi karakterlerin frekanslarını, 55'ten 57'ye kadar 3 nitelik büyük harflerle yazılmış kelimelerdeki toplam harf sayısı, ortalama harf sayısı ve en uzun kelimenin harf sayısını göstermektedir. 58. nitelik ise elektronik postanın istenmeyen (spam) olup olmadığını belirtmektedir. Toplam 4601 kayıttan 1813'ü istenmeyen elektronik posta (spam), 2788 elektronik posta ise normal elektronik postadır.

2. YÖNTEM

Veri madenciliğinde sınıflandırma eğiticili ve eğiticisiz olarak ikiye ayrılır. Eğiticili öğrenmede sınıf sayısı ve bir grup örneğin hangi sınıfa ait olduğu önceden bilinir. Hangi sınıfa ait olduğu bilinen kayıtlardan bir öğrenme kümesi model olarak alınır. Bu öğrenme kümesi modeli ile sınıfı bilinmeyen diğer kayıtlar denenerek hangi sınıfta olduğu bulunmaya çalışılır (Albayrak, 2015). Karar ağaçları eğiticili öğrenme sınıflandırmasında yaygın olarak kullanılan bir algoritma türüdür. Ayrıca karar ağaçları, eğer tahmin değeri kayıtların ait olduğu sınıftan geliyorsa sınıflandırma ağaçları veya tahmin değeri gerçek bir sayı kabul ediliyorsa (örneğin arabanın fiyatı) regresyon ağaçları olarak iki temel gruba ayrılabilir (L Breiman vd., 1984).

Sınıflandırma ağacı oluşturmak için öğrenme kümesindeki örnekleri en iyi belirleyen nitelik bulunur. Bu nitelik ile ağacın bir dalı ve yaprakları diye tabir edilen ayrıştırma yapılır ve yeni bir örnek kümesi oluşturulur. Ayrıştırılan bu dal üzerindeki örneklerden yeni bir belirleyici nitelik bulunur ve yeni dallar oluşturulur. Her bir alt veri kümesinde yani dal üzerindeki tüm örnekler aynı sınıfa aitse, örnekleri ayrıştıracak başka nitelik kalmamışsa ve kalan niteliklerinde değerini taşıyan başka örnek yoksa dallanma işlemi son bulur. Aksi halde alt veri kümesini ayrıştırmak için yeniden belirleyici bir nitelik bulunur (Albayrak, 2015).

Ayırt edici bu nitelik, bilgi kazancı (Cover & Thomas, 1991) veya gini indeksi hesaplamasıyla elde edilir. Hangi nitelik en büyük bilgi kazancını veriyorsa ağacın kökünde yer alacak özellik olarak o seçilir. Bilgi kazancını bulmak için entropi ölçümü kullanılır. Entropi rastgelelik, belirsizlik ve beklenmeyen bir durumun meydana gelme olasılığıdır. Bir sistemin bilgi kazancını hesaplamak için diyelim ki S, C_i sınıfından $i=(1, \dots, n)$ olmak üzere s_i kayıt içeriyorsa herhangi bir kaydı sınıflandırmak için gerekli bilgi denklem 1 ile hesaplanır.

$$Bilgi(s_1, s_2, \dots, s_n) = -\sum_{i=1}^{n} \frac{s_i}{S} \left(\log_2 \left(\frac{s_i}{S} \right) \right) \quad (1)$$

Bir A değişkeninin $(j=1,2,\dots,v)$ sayısı kadar alabildiği $(a_1, a_2, \dots, a_v)$ değerleri ile entropisini bulmak için ise denklem 2 kullanılır.

$$Entropi(A) = \sum_{j=1}^{v} \left(\frac{s_{1j} + \dots + s_{nj}}{S} \right) \cdot Bilgi(s_1, s_2, \dots, s_n) \quad (2)$$

A değişkenini kullanarak ağacın dallanmasıyla elde edilen bilgi kazancı denklem 3 ile bulunur.

$$Kazanç(A) = Bilgi(s_1, s_2, \dots, s_n) - Entropi(A) \quad (3)$$

Gini indeksi bütün değişkenlerin sürekli olduğunu kabul eder. Bu sürekliliği bozan en düşük gini indeks değerini veren ayrıma sahip değişken üzerinden bölünme gerçekleştirilir. Eğer bir T veri seti n farklı sınıfta N adet kayıt içeriyorsa p_i, j sınıfının T içindeki izafi sıklığını belirler ve Gini indeksi denklem 4 ile hesaplanır.

$$Gini(T) = 1 - \sum_{j=1}^{n} p_j^2 \quad (4)$$

Eğer T veri seti T_1 ve T_2 olarak sırasıyla N_1 ve N_2 büyüklüğünde ikiye ayrılan veri kümesi için gini indeksi, denklem 5'e göre en düşük değeri veren ayrıma sahip olan değişkenden seçilir.

$$Gini_{ayrim}(T) = \frac{N_1}{N} \cdot Gini(T_1) + \frac{N_2}{N} \cdot Gini(T_2) \quad (5)$$

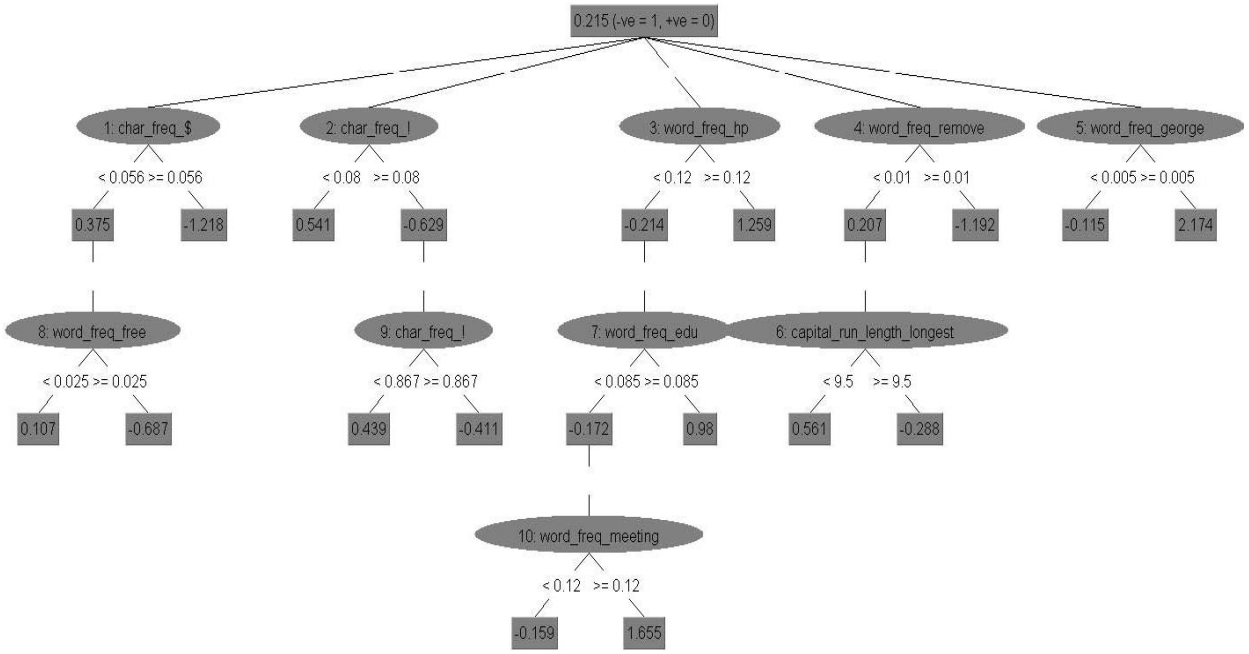
Sınıfların olmadığı veriler için ayrım kuralları, düğüm için tahmin edilen toplam varyansın azaltılması mantığına (L Breiman vd., 1984) dayanır ve böylece regresyon ağaçları oluşturulur.

3. UYGULAMA

İstenmeyen elektronik postaların *(spam)* tespiti için kullanılan veri seti Waikato Üniversitesi açık kaynak kodlu bir proje olan WEKA (Hall vd., 2009) makine öğrenmesi veri madenciliği yazılımı ile test edilmiştir. Bu çalışmada, 10 katlı çapraz doğrulama (Kohavi, 1995) tekniğine göre, eğitim ve test kümelerine ayrılan veriler üzerinde aşağıda açıklanan 12 farklı karar ağacı ile test edilmiştir.

3.1. Alternatif Karar Ağacı (Alternating Decision Tree – ADTree)

Alternatif Karar Ağacı (Alternating Decision Tree – ADTree) karar düğümleri ve tahmin düğümlerinden oluşur. Karar durumları bir eylem sonucu belirtir. Tahmin düğümleri tek bir sayı içerir. Alternatif karar ağaçları daima hem kök hem de yapraklar olarak tahmin düğümlerine sahiptir. Bir kayıt için sınıflandırma, üzerinden geçilen her tahmin düğümünün ve bütün karar düğümlerinin doğru olduğu yollar takip edilerek yapılır (Freund & Mason, 1999). C4.5 veya CART ağaçlarında ise bir kayıt sadece tek bir yoldan takip edilerek sınıflandırılır. İstenmeyen elektronik postaların ADTree ile sınıflandırılması için oluşturulan karar ağacı şekil 1'de gösterilmiştir.



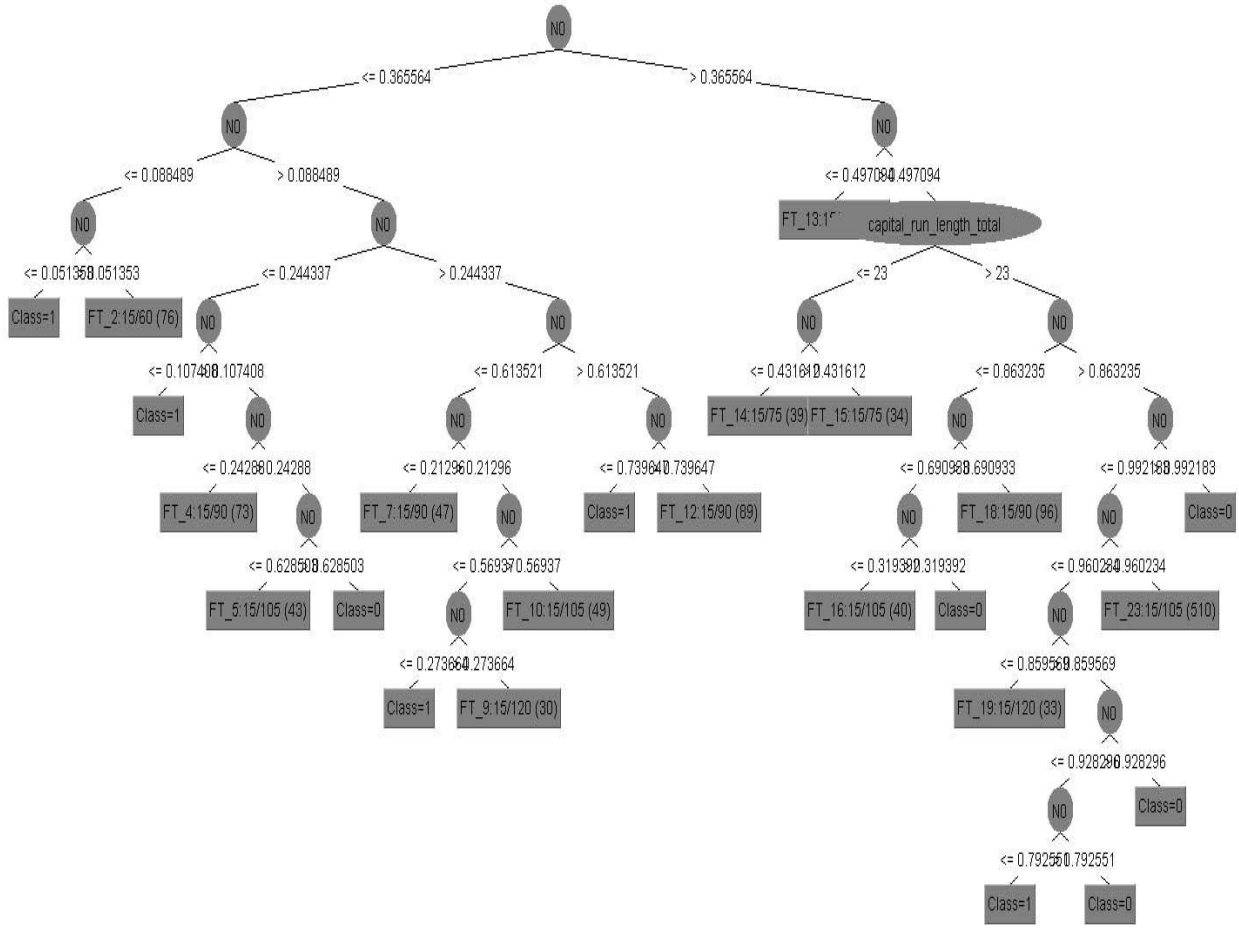
Şekil 1. ADTree ile istenmeyen elektronik posta *(spam)* sınıflandırması için karar ağacı

3.2. En iyi - ilk Karar Ağacı (Best-first Decision Tree – BFTree)

En iyi - ilk Karar Ağacı (Best-first Decision Tree – BFTree) böl ve yönet mantığına dayanır. Öncelikle bir nitelik kök olarak alınır ve bu nitelik üzerinden bazı kriterlere göre gruplara ayrılır. Daha sonra kök düğümden her grup genişletilerek eğitim verileri alt gruplara bölünür. Sadece karşılayan verileri kullanarak seçilen her grup için bu işlem tekrarlanır. Her işlemde, genişlemeye en uygun iyi alt grup seçilir (H. Shi, 2007). Bu süreç belirli bir genişleme katsayısına göre tüm düğümler net olana kadar devam eder.

3.3. Fonksiyonel Ağaçlar (Function Trees – FT)

Fonksiyonel Ağaçlar (Function Trees – FT) dört farklı model ile uygulanabilir. Fonksiyonel ağaçlar veriyi bir örnek uzayında birçok hiper-dikdörtgenlere bölerek sabitler. Giriş verilerinde karar uzayı, test edilen niteliğe dik ve diğer niteliklere paralel olmakla sorumludur. Bu sınıflandırıcılar tarafından oluşturulan bölge tamamen hiper-dikdörtgenlerdir. Fonksiyonel ağaçlar, bu biçimciliği nitelik kombinasyonlarına dayalı testlerin kullanımıyla genişletir (Gama, 2004). Tam fonksiyonel karar ağaçlarında karar düğümleri, nitelik kombinasyonu üzerine kurulu bir test verisi içerir ve yaprak düğümleri bir nitelik kombinasyonuna bağlı tahminlerde bulunur. Fonksiyonel ağaçlar ile sınıflandırılan istenmeyen elektronik postalar *(spam)* için oluşturulan karar ağacı şekil 2'de gösterilmiştir.



Şekil 2. FTree ile istenmeyen elektronik posta *(spam)* sınıflandırması için karar ağacı

3.4. J48 (C4.5) Karar Ağacı

J48 diğer adıyla C4.5 karar ağacı, en yüksek bilgi kazancına sahip nitelik üzerinden verilerin bölünmesiyle oluşturulan düğümlerden bir karar sonucuna ulaşır (Quinlan, 1993). Id3 algoritmasının bir türevidir.

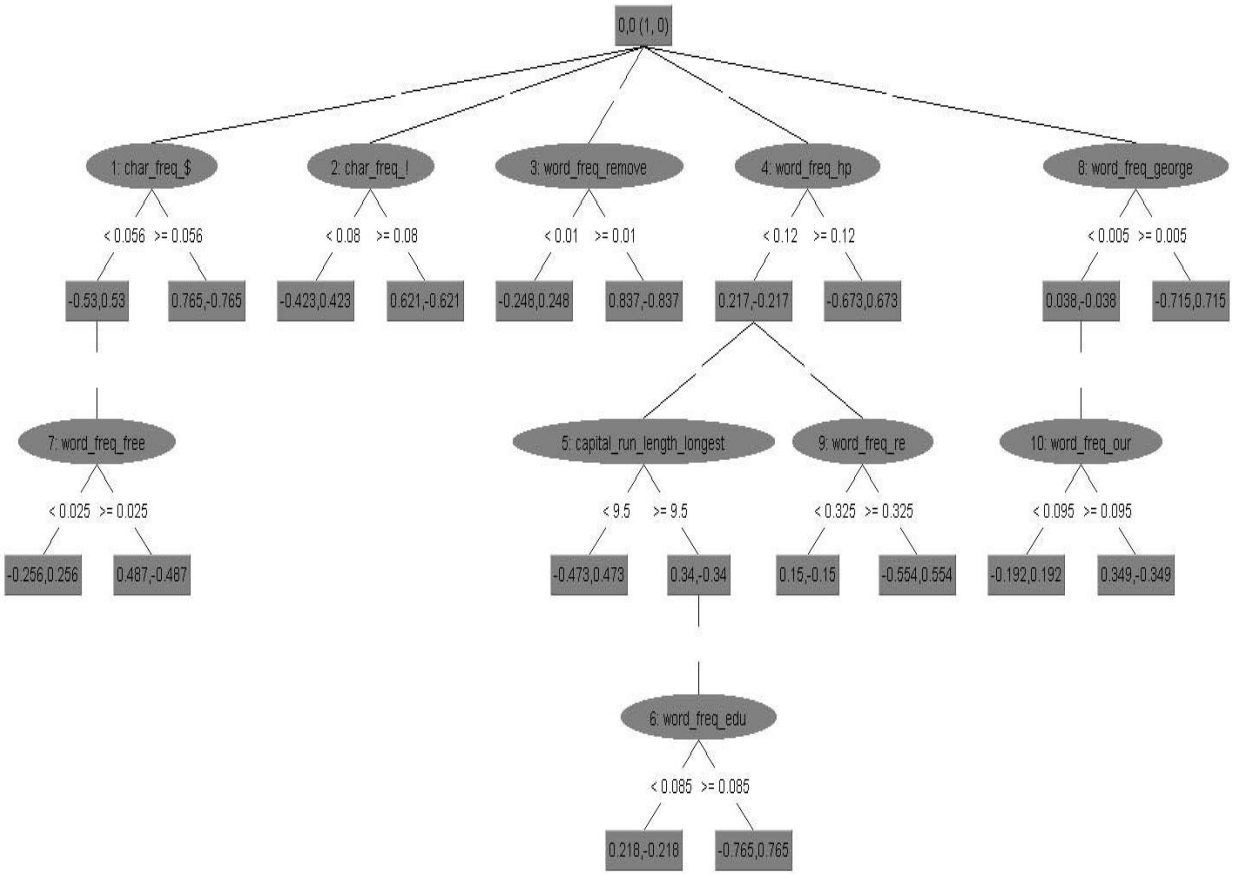
3.5. J48 Aşı (J48 graft) Karar Ağacı

J48 Aşı (J48 graft) karar ağacı, tahmin hatasını azaltmak amacıyla mevcut bir karar ağacına düğümler ekler. Yaprağa giden yolda en fazla bir testte başarısız olan olgular kümesinden, ilk karar ağacının her bir yaprakları için bir veri seti oluşturmayı amaçlayan bir algoritmadır. Bu teknik, ağacın karmaşıklığını ve oluşturma zamanını azaltırken, orijinal aşılama algoritmasının hata azaltma gücünü koruduğunu göstermiştir (Webb, 1999).

3.6. Verinin Mantıksal Analizi (Logical Analysis of Data – LADTree) Karar Ağacı

Verinin Mantıksal Analizi (Logical Analysis of Data – LADTree) karar ağacı, veri setindeki pozitif ve negatif örnekleri ayırt edebilen mantıksal bir ifadenin öğrenmesine dayalı ikili hedef değişkeni sınıflandırıcısıdır (Holmes vd., 2002). Girilen bir veri seti için model oluşumu, büyük grup desenleri oluşturmaya ve bunların içinden modeldeki her bir grubun

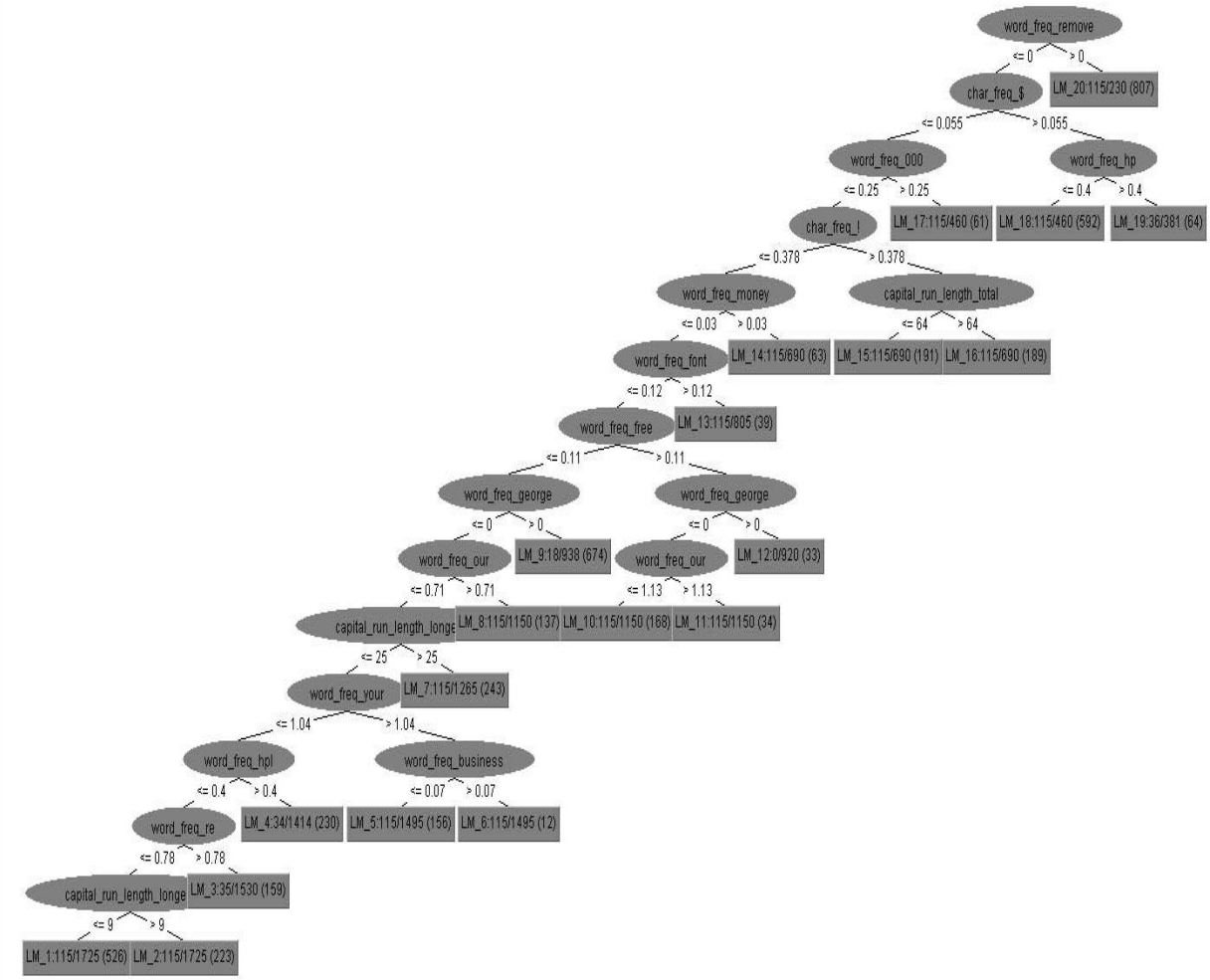
yaygınlık ve homojenlik açısından belirli gereksinimleri sağlamak şartıyla her biri yukarıda belirtilen varsayımları karşılayan alt grupları seçmeye dayanır (Wisaeng, 2013). LADTree ile istenmeyen elektronik posta *(spam)* tespitinde şekil 3 ile gösterilen karar ağacı kullanılmıştır.



Şekil 3. LADTree ile istenmeyen elektronik posta *(spam)* sınıflandırması için karar ağacı

3.7. Lojistik Modeli Ağacı (Logistics Model Tree – LMT) Algoritması

Lojistik Modeli Ağacı (Logistics Model Tree – LMT) algoritması, lojistik regresyon ve karar ağacını birleştiren öğrenimli bir eğitim sınıflandırma modelidir (Landwehr vd., 2003, 2005). Sıradan karar ağaçları yaprakları parçalı bir sabit modeli oluştururken, lojistik modeli ağacı yaprakları parçalı bir doğrusal regresyon sağlayan, doğrusal regresyon modeline sahip bir karar ağacıdır (Landwehr vd., 2003). Lojistik varyantta, LogitBoost algoritması, ağacın her düğümünde bir lojistik regresyon modeli üretmek için kullanılır; düğüm daha sonra C4.5 kriterleri kullanılarak ayrılır. Her LogitBoost yürütmesinde kendi sonuçlarından üst düğüm üzerinden yeniden başlatılır. Son olarak, ağaç budanır (Sumner vd., 2005). İstenmeyen elektronik posta tespitinde LADTree şekil 4 ile gösterilmiştir.



Şekil 4. LADTree ile istenmeyen elektronik posta *(spam)* sınıflandırması için karar ağacı

3.8. Naive Bayes (Naive Bayes Tree – NBTree) Karar Ağaçları

Naive Bayes (Naive Bayes Tree – NBTree) karar ağaçları, Bayes kuralı ile karar ağacını birleştiren öğreticili bir sınıflandırıcıdır. Bu algoritma (Kohavi, 1996), özellikleri, etiketi verilen koşullu bağımsız varsayarak, örneği verilen her sınıfın olasılığını hesaplamak için Bayes kuralını (Theodorescu, 1968) kullanır.

3.9. Rastgele Orman (Random Forest) Karar Ağacı

Rastgele Orman (Random Forest) karar ağacı, veri setinde en iyi niteliklerden seçilen düğümleri dallara ayırmak yerine, her bir düğümde rastgele alınan niteliklerin en iyisini seçerek tüm düğümleri dallara ayırır. Her veri kümesi asıl veri setinden yer değiştirmeli olarak üretilir. Rastgele özellik seçimi kullanılarak ağaçlar geliştirilir ve budama işlemi yoktur (Leo Breiman, 2001). Rastgele orman algoritmasının diğer algoritmalara göre daha hızlı ve doğru olmasının sebebi bu yöntemdir.

3.10. Rastgele Ağaç (Random Tree)

Rastgele Ağaç (Random Tree) her düğümde belli bir sayıda rastgele seçilmiş özellikleri alarak ağaç oluşturan bir sınıflama algoritmasıdır (Leo Breiman, 2001). Budama işlemi yoktur. Ayrıca tutulan veri setine dayalı sınıf olasılıklarının tahminine izin veren bir opsiyonu vardır.

3.11. REPTree Sınıflandırıcısı

REPTree sınıflandırıcısı, bölme kriteri olarak bilgi kazancını kullanarak ve azaltılmış budama hatası ile budama yaparak bir karar/regresyon ağacı oluşturan hızlı bir karar ağacı öğrenicisidir (Witten vd., 2011). Sadece sayısal verilerle çalışır ve kayıp veriler için C4.5 algoritması kullanır.

3.12. Basit Sınıflandırma ve Regresyon Ağacı (Simple Classification and Regression Tree – SimpleCART)

Basit Sınıflandırma ve Regresyon Ağacı (Simple Classification and Regression Tree – SimpleCART) algoritması sınıflandırma ve regresyon ağaçları birleşimi olan düğüm bölünmesinde benzerlikleri ve farklılıkları olan bir algoritmadır. Tahmin sonucu veriye ait olan sınıf ise bir sınıflandırma analizi, eğer tahmin sonucu örneğin bir hastanın hastane kalış süresi gibi gerçek bir sayı ise regresyon analizi olarak tanımlanır (L Breiman vd., 1984).

4. BULGULAR

İstenmeyen elektronik postaların tespiti için kullanılan veri seti Waikato Üniversitesi açık kaynak kodlu bir proje olan WEKA (Hall vd., 2009) makine öğrenmesi veri madenciliği yazılımı ile test edilmiştir. Bu çalışmada, 10 katlı çapraz doğrulama (Kohavi, 1995) tekniğine göre, eğitim ve test kümelerine ayrılan veriler üzerinde 12 farklı karar ağacı denenmiş ve elde edilen sonuçlar tablo 1'de gösterildiği gibi doğruluk oranları ve sınıflandırma modelinin oluşum zamanına göre kıyaslanmıştır.

Tablo 1. Karar ağaçları sınıflandırması sonuçları

Karar Ağacı	Doğru Sınıflandırma	Yanlış Sınıflandırma	Doğruluk Yüzdesi	Hata Yüzdesi	Model Oluşum Zamanı
Random Forest	4356	245	94.6751	5.3249	2.11 saniye
Logistic Model Tree	4313	288	93.7405	6.2595	419.5 saniye
J48 (C4.5) Grafted	4292	309	93.2841	6.7159	2.53 saniye
Functional Tree	4290	311	93.2406	6.7594	7.42 saniye
Naive Bayes Tree	4288	313	93.1971	6.8029	185.76 saniye
J48 (C.5)	4278	323	92.9798	7.0202	1.65 saniye
Reduced Error Pruning Tree	4270	331	92.8059	7.1941	1.34 saniye
Best-First Decision Tree	4267	334	92.7407	7.2593	6.96 saniye
Simple CART	4245	356	92.2626	7.7374	10.02 saniye
Alternating Decision Tree	4239	362	92.1321	7.8679	5.69 saniye
LogitBoost Alternating Decision	4237	364	92.0887	7.9113	11.16 saniye
Random Tree	4189	412	91.0454	8.9546	0.15 saniye

Elde edilen sonuçlara göre Random Forest algoritması kullanılarak, 2.11 saniyede 4601 adet elektronik postadan %94.68 doğruluk oranı ile istenmeyen elektronik postaları *(spam)* tespit etmiştir. Ancak diğer algoritmalarla kıyaslandığında, sınıflandırma modelinin oluşum

zamanı daha kısa olan algoritmalar mevcuttur. Örneğin, çok bilinen karar ağacı yöntemlerinden J48 (C4.5) algoritması 1.65 saniyede %93 doğruluk oranı yakalamıştır. Ayrıca sonuç tablosundan anlaşılacağı üzere algoritmaların doğruluk oranlarının birbirine yakın olduğu görülmektedir. Bununla beraber, her ne kadar doğruluk oranı yüksek olsa da Naive Bayes veya Logistic Model karar ağaçları, model oluşum süresi bakımından kullanışlı olmadığı tespit edilmiştir.

Bu sınıflandırma testinin yeterliliğini değerlendirmek için alıcı işlem karakteristiği (Reciever Operating Characteristics) ROC analizi (Gribskov & Robinson, 1996) yapılarak tablo 2'de gösterilen durumlara göre hassasiyet ve kesinlik ölçütleri hesaplanmıştır. Sınıflandırma işleminde metotlar, doğru pozitif değerleri (hassasiyet) tespit etme ve yanlış pozitif değerleri (kesinlik) eleme kabiliyeti arasındaki dengeyi kurmakla uğraşırlar. Kesinlik ve hassasiyet arasındaki dengeyi değerlendirmek için ROC eğrisi ve altında kalan alan kullanılır. ROC puanı 1'e yaklaştıkça pozitifler daha iyi bir şekilde negatiflerden ayrılır.

Tablo 2. ROC analizi sonuçları

Karar Ağacı	DP	DN	YP	YN	Hassasiyet	Kesinlik	ROC puanı
Random Forest	1696	2660	117	128	0.930	0.958	0.982
Alternating Decision	1612	2627	201	161	0.909	0.929	0.972
LogitBoost Alternating Decision	1620	2617	193	171	0.905	0.931	0.969
Naive Bayes	1657	2631	156	157	0.913	0.944	0.968
Logistic Model	1660	2653	153	135	0.925	0.945	0.967
Reduced Error Pruning	1637	2633	176	155	0.914	0.937	0.959
Functional	1673	2617	140	171	0.907	0.949	0.949
Simple CART	1616	2629	197	159	0.910	0.930	0.945
C4.5 grafted	1640	2652	173	136	0.923	0.939	0.941
C4.5	1646	2632	167	156	0.913	0.940	0.939
Best-First Decision	1622	2645	191	143	0.919	0.933	0.936
Random Tree	1627	2562	186	226	0.878	0.932	0.908

DP (doğru pozitif), gerçekte spam olan ve algoritmanın da spam tespit ettiği elektronik postalar
DN (doğru negatif), gerçekte spam olmayan ve algoritmanın spam tespit etmediği elektronik postalar
YP (yanlış pozitif), gerçekte spam olan ama algoritmanın spam tespit etmediği elektronik postalar
YN (yanlış negatif), gerçekte spam olmayan ama algoritmanın spam tespit ettiği elektronik postalar
Hassasiyet DP/(DP+YN) oranı
Kesinlik DN/(DN+YP) oranı

ROC analizi sonuçlarına göre Random Forest 0.958 ROC puanı ile 1'e en yakın algoritma olarak tespit edilmiştir. Algoritmanın hassasiyet ve kesinlik açısından da diğerlerine göre daha başarılı olduğu anlaşılmaktadır.

Spambase veri seti üzerinde, 10 katlı çapraz doğrulama ile yapılan sınıflandırma analizinde Random Forest karar ağacı %94.68 en iyi sınıflandırma başarısını elde etmiştir. Aynı veri setini kullanan, kaba kümeler yaklaşımı çalışmasında (Kaya vd., 2011) %91.23 doğruluk oranı elde edilirken, Aşırı Öğrenme Makinesi (Extreme Learning Machine) algoritması ile yapılan çalışmada (Kaya vd., 2014) %91.66 sınıflandırma başarı elde edilmiştir. Aynı veri setini kullanan diğer bir çalışmada (Kumar vd., 2012) ise farklı makine öğrenmesi yöntemlerinden yapay sinir ağları, doğrusal diskriminant analizi, K yakın komşuluk algoritması, regresyon analizleri ve karar ağaçları karşılaştırması yapılmıştır.

5. SONUÇ

İnternet kullanımının artmasıyla beraber, elektronik posta kullanımı da yaygınlaşmıştır. Ancak beraberinde getirdiği sorunlardan, istenmeyen elektronik postaların *(spam)* tespit edilmesi de bir zorunluluk halini almıştır. Bu çalışmada, istenmeyen elektronik posta *(spam)* tespiti için 12 farklı karar ağacı uygulamasının performansları test edilmiş ve en uygun karar ağacı türü belirlenmiştir.

Kullanılan veri setinde, elektronik postaların içinde geçen kelimelerin ve karakterlerin frekanslarına göre karar ağaçları ile istenmeyen elektronik postalar *(spam)* tespit edilmiştir. Karar ağacı sınıflandırıcıların performans ölçümü için ROC analizi yapılarak doğruluk, hassasiyet ve kesinlik değerleri ile ROC puanı hesaplanmıştır. Ayrıca elektronik postaların alıcıya ulaşmadan önce filtrelenmesi gerektiği için, sınıflandırma algoritmalarının model oluşum zamanları dikkate alınmış ve yoğun trafiğe sahip elektronik posta alışverişi için en uygun yöntemin hangisi olacağı da analiz edilmiştir. Aynı veri setini kullanan diğer çalışmalara oranla karar ağaçları daha iyi bir performans göstermiştir. Karar ağaçlarının spam filtreleme için doğruluk ve algoritma hızı açısından başarılı bir veri madenciliği yöntemi olduğu tespit edilmiştir.

Elde edilen sonuçlara göre diğer yöntemler içinde en iyi doğruluk oranı ve zamanlama başarısını gösteren Random Forest karar ağacı, geniş kapasiteli veri setlerini hızlı ve efektif bir şekilde sınıflandırabildiği için bu algoritmanın bu tür çalışmalarda, diğer algoritmalara kıyasla daha kullanışlı bir algoritma olduğu tespit edilmiştir.

Karar ağaçlarının farklı yöntemler kullanılarak örneğin optimizasyon algoritmaları ile melezleştirilmesi sayesinde performans artırımı ve/veya çok dilli elektronik postaların sınıflandırılması ileriye dönük çalışmalar açısından önemli sonuçlar ortaya koyabilecek araştırma konuları olacaktır.

KAYNAKÇA

Ahmed, F., & Abulaish, M. (2013). A generic statistical approach for spam detection in Online Social Networks. *Computer Communications, 36*(10–11), 1120-1129. doi: http://dx.doi.org/10.1016/j.comcom.2013.04.004

Albayrak, S. (2015). *CE 4850 Data Mining*. Sınıflama ve Kümeleme Yöntemleri. Ders Notları. Bilgisayar Mühendisliği. Yıldız Teknik Üniversitesi. Istanbul. Retrieved from http://www.ce.yildiz.edu.tr/personal/songul/file/332/Veri%20Madencili%C4%9Fi-S%C4%B1n%C4%B1flamaKumeleme.ppt

Bache, K., & Lichman, M. (2013). *UCI Machine Learning Repository*. Retrieved from: http://archive.ics.uci.edu/ml

Breiman, L. (2001). Random Forests. *Machine Learning, 45*(1), 5-32. doi: 10.1023/A:1010933404324

Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (1984). *Classification and regression trees*. Monterey, CA: Wadsworth & Brooks/Cole Advanced Books & Software.

Chan, P. P. K., Yang, C., Yeung, D. S., & Ng, W. W. Y. (2015). Spam filtering for short messages in adversarial environment. *Neurocomputing, 155*(0), 167-176. doi: http://dx.doi.org/10.1016/j.neucom.2014.12.034

Cover, T. M., & Thomas, J. A. (1991). *Elements of Information Theory*. New York: Wiley & Sons.

El-Alfy, E. S. M. (2009, 18-21 May 2009). *Discovering classification rules for email spam filtering with an ant colony optimization algorithm.* Paper presented at the Evolutionary Computation, 2009. CEC '09. IEEE Congress on.

Fawcett, T. (2003). ``*In vivo'' Spam Filtering: A challenge problem for data mining.* Paper presented at the ACM SIGKDD Explorations Newsletter.

Freund, Y., & Mason, L. (1999). *The Alternating Decision Tree Learning Algorithm*. Paper presented at the Proceedings of the Sixteenth International Conference on Machine Learning.

Gaikwad, B. U., & Halkarnikar, P. Spam E-mail Detection by Random Forests Algorithm.

Gama, J. (2004). Functional Trees. *Machine Learning, 55*(3), 219-250. doi: 10.1023/B:MACH.0000027782.67192.13

Gribskov, M., & Robinson, N. L. (1996). Use of receiver operating characteristic (ROC) analysis to evaluate sequence matching. *Computers & Chemistry, 20*(1), 25-33. doi: http://dx.doi.org/10.1016/S0097-8485(96)80004-0

Guzella, T. S., & Caminhas, W. M. (2009). A review of machine learning approaches to Spam filtering. *Expert Systems with Applications, 36*(7), 10206-10222. doi: http://dx.doi.org/10.1016/j.eswa.2009.02.037

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., & Witten, I. H. (2009). The WEKA data mining software: an update. *SIGKDD Explor. Newsl., 11*(1), 10-18. doi: 10.1145/1656274.1656278

Holmes, G., Pfahringer, B., Kirkby, R., Frank, E., & Hall, M. (2002). *Multiclass Alternating Decision Trees*. Paper presented at the Proceedings of the 13th European Conference on Machine Learning.

Kaya, Y., Ertuğrul, Ö. F., & Tekin, R. (2014). An Expert Spam Detection System Based on Extreme Learning Machine. *Computer Science and Applications, 1*(2), 132-137.

Kaya, Y., Yeşilova, A., & Tekin, R. (2011, 5-6-7 Eylül 2011). *İstenmeyen Elektronik Postaların(Spam) Filtrelenmesinde Kaba Küme Yaklaşımının Kullanılması.* Paper presented at the Elektrik-Elektronik Bilgisayar Sempozyumu (FEEB 2011), Elazığ TURKEY.

Kohavi, R. (1995). *A study of cross-validation and bootstrap for accuracy estimation and model selection*. Paper presented at the Proceedings of the 14th international joint conference on Artificial intelligence - Volume 2, Montreal, Quebec, Canada.

Kohavi, R. (1996). *Scaling Up the Accuracy of Naive-Bayes Classifiers: A Decision-Tree Hybrid.* Paper presented at the KDD.

Kumar, R. K., Poonkuzhali, G., & Sudhakar, P. (2012, March 14-16). *Comparative Study on Email Spam Classifier using Data Mining Techniques* Paper presented at the IMECS 2012 : International MultiConference of Engineers and Computer Scientists Hong Kong.

Landwehr, N., Hall, M., & Frank, E. (2003). Logistic Model Trees. In N. Lavrač, D. Gamberger, H. Blockeel & L. Todorovski (Eds.), *Machine Learning: ECML 2003* (Vol. 2837, pp. 241-252): Springer Berlin Heidelberg.

Landwehr, N., Hall, M., & Frank, E. (2005). Logistic Model Trees. *Mach. Learn., 59*(1-2), 161-205. doi: 10.1007/s10994-005-0466-3

Laorden, C., Santos, I., Sanz, B., Alvarez, G., & Bringas, P. G. (2012). Word sense disambiguation for spam filtering. *Electronic Commerce Research and Applications, 11*(3), 290-298. doi: http://dx.doi.org/10.1016/j.elerap.2011.11.004

Orhan, U. (2012). *Makine Öğrenmesi Dersi*. Entropi, Karar Ağaçları (ID3 ve C4.5 algoritmaları), Sınıflandırma ve Regresyon Ağaçları. Ders Notları. Müh.Mim. Fakültesi Bilgisayar Mühendisliği. Çukurova Universitesi. Adana. Retrieved from http://bmb.cu.edu.tr/uorhan/DersNotu/Ders03.pdf

Quinlan, J. R. (1993). *C4.5: programs for machine learning*: Morgan Kaufmann Publishers Inc.

Ruano-Ordás, D., Fdez-Glez, J., Fdez-Riverola, F., & Méndez, J. R. (2013). Effective scheduling strategies for boosting performance on rule-based spam filtering frameworks. *Journal of Systems and Software, 86*(12), 3151-3161. doi: http://dx.doi.org/10.1016/j.jss.2013.07.036

Shi, H. (2007). *Best-first Decision Tree Learning.* (Master od Science (MSc)), The University of Waikato, Hamilton, NewZealand. Retrieved from http://researchcommons.waikato.ac.nz/bitstream/handle/10289/2317/thesis.pdf

Shi, L., Wang, Q., Ma, X., Weng, M., & Qiao, H. (2012). Spam email classification using decision tree ensemble. *Journal of Computational Information Systems, 8*(3), 949-956.

Sumner, M., Frank, E., & Hall, M. (2005). Speeding Up Logistic Model Tree Induction. In A. Jorge, L. Torgo, P. Brazdil, R. Camacho & J. Gama (Eds.), *Knowledge Discovery in Databases: PKDD 2005* (Vol. 3721, pp. 675-683): Springer Berlin Heidelberg.

Theodorescu, R. (1968). Good, I. J.: The Estimation of Probabilities. An essay on modern Bayesian Methods. Research Monograph No. 30, The M. I. T. Press Cambridge 1965. 109 S., 7 Tab., 115 Lit. *Biometrische Zeitschrift, 10*(1), 87-87. doi: 10.1002/bimj.19680100118

Webb, G. I. (1999). *Decision tree grafting from the all-tests-but-one partition*. Paper presented at the Proceedings of the 16th international joint conference on Artificial intelligence - Volume 2, Stockholm, Sweden.

Wisaeng, K. (2013). A Comparison of Different Classification Techniques for Bank Direct Marketing. *International Journal of Soft Computing and Engineering (IJSCE), 3*(4), 116-119.

Witten, I. H., Frank, E., & Hall, M. A. (2011). *Data Mining: Practical Machine Learning Tools and Techniques: Practical Machine Learning Tools and Techniques*: Elsevier Science.

Wu, C.-H., & Tsai, C.-H. (2009). Robust classification for spam filtering by back-propagation neural networks using behavior-based features. *Applied Intelligence, 31*(2), 107-121. doi: 10.1007/s10489-008-0116-0

Ying, K.-C., Lin, S.-W., Lee, Z.-J., & Lin, Y.-T. (2010). An ensemble approach applied to classify spam e-mails. *Expert Systems with Applications, 37*(3), 2197-2201. doi: http://dx.doi.org/10.1016/j.eswa.2009.07.080