Gerçek Veri Setlerinde Klasik Makine Öğrenmesi Yöntemlerinin Performans Analizi

Metin BİLGİN[1]

[1] Bursa Teknik Üniversitesi, DBMM Fakültesi, BURSA

Özet : Bu çalışmada, Biyomedikal veri setleri üzerinde Klasik Makine Öğrenmesi (KMÖ) yöntemlerinin sınıflandırma performansları analiz edilmiştir. Yapılan çalışmada UCI biyomedikal veri seti üzerinden 6 farklı veri seti edinilmiş ve WEKA programı içerisinde bulunan 6 farklı makine öğrenmesi ile sistem çalıştırılmıştır. Çapraz doğrulama yöntemi k=10 olacak şekilde kullanılmıştır. Yapılan çalışmalar sonucunda Klasik Makine Öğrenmesi yöntemleri içinde Sequential Minimal Optimization (Sıralı Minimal Optimizasyon-SMO) sınıflandırma doğruluğu için diğer beş yönteme göre daha yüksek performans göstermiştir. Yapılan çalışma ile literatüre yeni bir yöntem önerilmemiştir. Gerçekleştirilen çalışma ile bir araştırmanın sonuçları aktarılmıştır. Her ne kadar bu çalışma sonunda yeni bir yöntem önerisi getirilmese de bu çalışmalardan elde edilen veriler ışında üzerinde çalışılan yeni bir makine öğrenmesi yöntemi için temel karşılaştırma parametrelerini vermektedir.

Anahtar Sözcükler: Makine Öğrenmesi, Biyomedikal Veri, Weka

Abstract :
Keywords:

1.Giriş

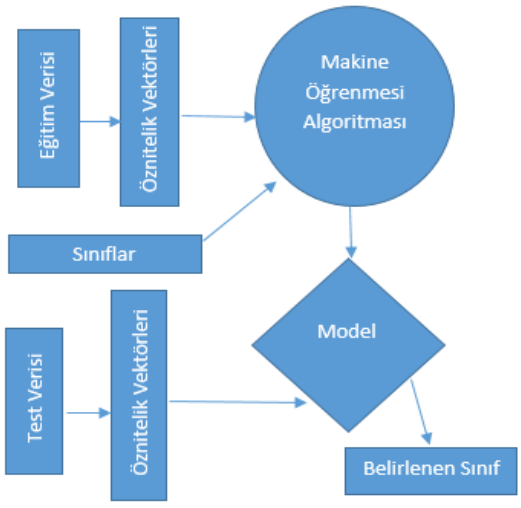
Makine Öğrenmesi (MÖ), önceki gözlemlere dayanarak doğru tahminler yapabilmeyi öğrenebilmek amacıyla otomatik tekniklerin geliştirilmesidir [1]. Ayodele'ye göre ise MÖ, otomatik olarak öğrenme işlemini deneyimlerden yola çıkarak geliştiren ve gerçekleştiren bilgisayar sistemlerinin geliştirilmesidir [2].

MÖ üzerine yapılan araştırmaların artmasıyla birlikte hesaplamalı öğrenme teorisi, yapay sinir ağları, istatistik ve örüntü tanıma gibi araştırma alanları arasında bağlantı kurulmuş ve bu alanlar birlikte çalışılabilmiştir. Böylece MÖ teknikleri daha geleneksel problemlerin (örneğin yüz tanıma) yanı sıra veritabanlarında bilgi keşfi, dil işleme ve robot kontrolü gibi yeni problemlere uygulanmaya başlamıştır [3].

MÖ'de veri çok önemli bir rol oynamakta ve öğrenme algoritmaları veriye ait bilgi ve özelliklerin keşfedilmesinde kullanılır [4]. Kullanılan veri ise etiketli ve etiketsiz olmak üzere iki ayrı türden oluşmaktadır. Etiketli veri seti bir algoritmayı eğitmek, etiketsiz veri ise eğitilmiş algoritmayı (model veya sistemi) test etmek için kullanılmaktadır. Bu nedenle bu veri tipleri eğitim ve test seti olarak da adlandırılmaktadır. MÖ kullanılarak oluşturulan sistemler genelde iki farklı öğrenme modeli kullanmaktadır. Bu iki model denetimli (supervised) ve denetimsiz (unsupervised) öğrenme modeli olarak adlandırılmaktadır [5].

Bunun yanı sıra her iki modelin birlikte kullanıldığı öğrenme sistemleri (örneğin semisupervised learning) de bulunmaktadır.

Denetimli öğrenme Denetimli Makine Öğrenmesi (SML: Supervised Machine Learning) sistemin etiketli veriler kullanılarak eğitilmesi ile öğrenmenin sağlanmasıdır. Sistem eğitilirken veri setinde bulunan her bir örneğe ait giriş ve çıkışlar verilir. Metin Sınıflandırma çalışmalarında giriş metnin içeriğini, çıkış ise kategorisini temsil eder. Test veri seti ise sistemin doğrulanması amacıyla kullanılır. Sistemin doğrulanması aşamasında öğrenme algoritması kategorisi bilinmeyen bir test verisine, eğitim verisinde bulunan çıkışlardan herhangi birini atar [6]. Denetimli öğrenme modeli süreci Şekil 1'de verildiği gibi gerçekleşmektedir [7].



Şekil 1 Denetimli Öğrenme Modeli Süreci

Denetimli öğrenme modelinde problem, sınıflandırma problemi olarak ele alınır ve eğitilmiş sistem test setine yönelik tahmin ve tanıma amacıyla kullanılır [4]. Destek Vektör Makinesi, Yapay Sinir Ağları, Lojistik Regresyon, Basit Bayes, Multinom Basit Bayes, k-En Yakın Komşu, Rastgele Orman ve Karar Ağaçları algoritmaları denetimli öğrenme modeli oluşturulurken yaygın olarak kullanılan yöntemlerdendir [8].

Sınıflandırma problemlerinde kullanılan yöntemler de uygulanan problem veya sonuç beklentisindeki farklılığa göre değişiklik gösterebilmektedir. Bu nedenle sınıflandırma sonuçlarının tek etiketli veya çok etiketli olarak üretilmesi gerekebilmektedir. Tek etiketli sınıflandırmada (single-label) test örneğine mevcut etiket kümesinden sadece bir etiket atanırken, çok etiketli sınıflandırmada (multi-label) birden fazla etiket atanmaktadır. Ayrıca tek etiketli sınıflandırmanın özel bir durumu olan ikili sınıflandırmada (binary classification) ise mevcut iki kategori arasından tek bir etiket atanarak sınıflandırma yapılmaktadır [9].

Denetimsiz öğrenme Denetimsiz Makine Öğrenmesi (UML: Unsupervised Machine Learning) modelinde sistem eğitilirken etiketsiz veri kullanılarak öğrenmesi sağlanır. Denetimsiz öğrenmede amaç veri setindeki örneklerin çıkışları bilinmediği için tanıma veya sınıflandırma değildir. Genellikle kümeleme, olasılık yoğunluk tahmini, öznitelikler arasındaki ilişkilerin bulunması ve boyut indirgeme gibi amaçlarla kullanılmaktadır. Ayrıca denetimsiz öğrenme algoritması ile elde edilen sonuçlar denetimli öğrenme için de kullanılabilmektedir [4]. Parçalayıcı ve hiyerarşik kümeleme algoritmaları ise genellikle denetimsiz öğrenme modeli oluşturulurken kullanılan algoritmalardır [10].

2.Yapılandırma

2.1. WEKA

WEKA, bilgisayar bilimlerinin önemli konularından birisi olan MÖ konusunda kullanılan paketlerden birisinin ismidir. Waikato üniversitesinde açık kaynak kodlu olarak JAVA dili üzerinde geliştirilmiştir ve GPL lisansı ile dağıtılmaktadır. İsmi de buradan gelir ve Waikato Environment for Knowledge Analysis kelimelerinin baş harflerinden oluşur.

WEKA verileri basit bir dosyadan okur ve veriler üzerindeki skolastik değişkenlerin sayısal veya nominal değerler olduğunu kabul eder. Aynı zamanda veri tabanı (database) üzerinden de veri çekebilir ancak bu durumda verilerin bir dosya verisi şeklinde olması beklenir.

WEKA üzerinde makine öğrenmesi ve istatistik ile ilgili pek çok kütüphane hazır olarak gelmektedir. Örneğin veri ön işlemesi (data preprocessing), ilkelleme (regression), sınıflandırma (classification), gruplandırma (clustering), özellik seçimi veya özellik çıkarımı (feature extraction) bunlardan bazılarıdır. Ayrıca bu işlemler sonucunda çıkan neticelerinde görsel olarak gösterilmesini sağlayan görüntüleme (visualization) araçları bulunmaktadır [**12].**

2.2. Arff Dosya Formatı

İngilizce, Attribute Relationship File Format kelimelerinin baş harflerinden oluşmuştur. ARFF dosya yapısı, Weka'ya özel olarak geliştirilmiştir ve dosya, metin yapısında tutulmaktadır. Dosyanın ilk satırında, dosyadaki ilişki tipi (relation) tutulmakta olup ikinci satırdan itibaren veri kümesindeki özellikler (attributes) yazılmaktadır. Özelliklerin hemen ardından veri kümesi yer alır ve veri kümesindeki her satır bir örneğe (instance) işaret etmektedir. Ayrıca veri kümesindeki her örneğin her özelliği arasında da virgül ayıracı kullanılmaktadır. Arff formatında yazılmış örnek kod Şekil 2'de görülmektedir.

```
@relation havatahmini

@attribute nem numeric
@attribute sıcaklık numeric
@attribute basınç numeric
@attribute tahmin numeric

@data
53,25,1013,1
41,22,1011,-1
54,18,1012,-1
67,23,1000,1
```

Şekil 2 Arff Dosya Örneği

Şekil 2'de verilen örnek dosyada, hava tahmini için kullanılan nem, sıcaklık ve basınç değerleri bir dosya içerisinde 4 örnek içerecek şekilde gösterilmiştir. Bu değerler tip olarak sayısal değerler olduğundan "numeric" olarak ifade edilmiştir. Ancak bu değerler aşağıdaki tiplerde olabilir:

NOMINAL: [Küme Değerleri] Tahmin değeridir ve bir tanım kümesi alır. Örneğin tahmin {güneşli,yağmurlu,sisli} şeklinde tanımlanan bir kümede, bu özellik kümedeki tanımlı değerlerden birisini alabilir.

REAL: [Reel Sayılar] kümesinden bir değer verileceğinde kullanılır. Örneğin sıcaklık değeri 22.8 şeklinde ondalıklı değerleri de ifade edecek şekilde verilmek istenirse tip olarak numeric yerine reel kullanabiliriz.

STRING: Veri kümesinin bu özelliğinin serbest yazı şeklinde olabileceğini ifade eder. Özellikle metin madenciliği çalışmaları için sıkça kullanılan bir tiptir.

DATE: Veri kümesinin bu özelliğinin tarih olduğunu ifade eder. Örneğin veri kümesindeki kişilerin doğum tarihi veya örneklerin toplanma tarihi gibi özelliklerin tutulmasında kullanılabilir [**13]**

2.3.Veri Kümesi

Bu çalışmada UCI veri kümesinden elde edilen 6 farklı veri kümesine ait bilgiler Çizelge 1'de verilmiştir.

No	Veri Kümesi	Sınıf Sayısı	Özellik Sayısı	Örnek Sayısı
1	Pima	2	8	768
2	Breast	2	9	683
3	Spect Heart	2	22	267
4	Dermatology	6	34	358
5	Ecoli	8	8	336
6	Yeast	10	9	1484

Çizelge 1 Kullanılan Veri Kümesine ait bilgiler

Pima: Hastanın hamilelik sayısı, yaşı, öz sıvıdaki insülin miktarı, glükoz yoğunluğu, derideki kıvrım kalınlığı ve vücut kitle indeksi gibi değerleri özellik olarak kullanan bir veri kümesidir.

Breast Cancer: Hastalık bölgesindeki tümörlü alanın özellikleri sınıflandırmada nitelik olarak kullanılmaktadır. Örneğin: tümörlü bölgenin kalınlığı, tümörlü hücrelerin şekil ve büyüklük olarak birbirine benzemesi oranı, tümörün vücuda yapışma oranı gibi özellikler. Bu özelliklerin kullanılması ile tümörün iyi huylu veya kötü huylu olması sonucu elde edilir.

Spect Heart, ek foton yayan radyonüklid'le işaretlenmiş doğal biyokimyasal madde (glükoz vb.) enjeksiyonunu takiben, incelenen organdaki gama ışınları dağılımının, hasta çevresinde dairesel dönüş gösteren dedektörlerle belirlenerek ilgili organın normal yada anormal olduğunun tespitidir.

Ecoli, Genelde E. coli kısaltması ile veya koli basili olarak bilinen Escherichia coli, memeli hayvanların kalın bağırsağında yaşayan bakteri türlerinden biridir.

Dermatology, Ciltte oluşan kızarıklık, sertleşme, kaşıntı, yumuşama, kepeklenme, çatlama; kafa derisi, diz ve el dirseklerinde hastalığın görülme düzeyi gibi medikal özellikler kullanılarak hastanın hastalık sınıfı tahmin edilmektedir. Hastalık sınıfları: Sedef Hastalığı, Seboreik Egzama, Liken Planus, Pitriyazis Rozea, Kronik Deri İltihabı, Pitriyazis Rubra Pilar.[15]

Yeast, Mantar enfeksiyonları bir veya birkaç mantar türüyle dokuların istilasını temsil eder. Yüzeyel, yerel doku hastalıkları ile daha derin, ciddi akciğer, kan (septisemi) veya sistemik hastalıkları içeren çok sayıda mantar enfeksiyonu vardır. Bazı mantarlar fırsatçı iken diğerleri hastalık etkeni (patojenik) olup bağışıklık sistemi sağlıklı olsun olmasın hastalığa neden olurlar.

3. Deney

Gerçekleştirilen çalışmada 6 farklı veri kümesine ait veriler 10 parça çapraz doğrulama (Cross Validation) yöntemi ile eğitilip ardından test edilmiştir. Çapraz doğrulamada veri kümesi öncelikle belirlenen alt küme sayısına bölünür. Alt küme sayısı bu çalışmada 10 seçilmiştir. Ardından her parça 1 kez test kümesi olacak şekilde sistem eğitilir ve ardından test kümesi ile test edilir. Çıkan test sonuçlarının ortalaması alınarak sistemin başarısı ölçülür.

Makine öğrenmesi yöntemleri olarak Naive Bayes (NB), Random Forest (Rastgele Orman-RF), Sequential Minimal Optimization (Sıralı Minimal Optimizasyon-SMO), IBk (K-en yakın komşu, k-nearest neigbour), Decision Table (Karar Tabloları-DT), J48 kullanılmıştır.

3.1. Kullanılan Makine Öğrenmesi Yöntemleri

NB, Naive Bayes Sınıflandırıcısı Bayes teoremine dayanan basit bir olasılıksal sınıflandırma yöntemidir. Mevcut sınıflanmış durumdaki örnek verileri kullanarak yeni bir verinin mevcut sınıflardan herhangi birine ait olma olasılığını hesaplayan bir yaklaşımdır. Bu sınıflandırıcıda

nitelikler birbirinden bağımsız olarak kabul edilir. [15]

RF, Breiman tek bir karar ağacı üretmek yerine çok sayıda ve çok değişkenli ağaçların her birinin farklı eğitim kümeleriyle eğitilmesi sonucu ortaya çıkan kararların birleştirilmesini önerir. Bir sınıflandırıcı yerine birden çok sınıflandırıcı üreten ve sonrasında onların tahminlerinden alınan oylar ile yeni veriyi sınıflandıran öğrenme algoritmasıdır. Büyük veri tabanlarında eşsiz olarak çalışır ve dengesiz veri kümesi sınıfında hata dengeleme yöntemlerine sahiptir. Kaybolan verilerin büyük olasılığında doğruluk korunur ve kaybolan verilerin tahmin edilmesinde etkili bir metottur [16][17]

SMO, herhangi bir ekstra matris depolama olmadan ve tüm sayısal QP (Quatratic Programming) optimizasyon adımları kullanmadan SVM QP sorununu hızlı bir şekilde çözer [18]. Bu uygulama global olarak bütün kayıp değerleri yenisiyle değiştirir ve nominal öznitelikleri ikili olanlara dönüştürür. Ayrıca bütün öznitelikleri (attributes) önceden tanımlanmış değerlerle (default) normalize eder.

IBk, algoritması öznitelik uzayındaki en yakın eğitim örneklerine dayanarak nesneleri sınıflandıran, en basit örüntü tanıma yöntemlerinden birisidir. Bu algoritma verilen k değeri kadar en yakın komşunun sınıfına göre sınıflandırma işlemi yapmaktadır. IBk algoritmasında bir vektörün sınıflandırılması, sınıfı bilinen vektörler kullanılarak yapılmaktadır.

DT, algoritma sınıflandırma için bir karar tablosu oluşturur. Eğitim verilerinin öz niteliklerine göre ortaya çıkan bu karar tablosundan faydalanarak sınıflandırma yapar [19].

J48, J. Ross Quinlan tarafından geliştirilen çok popüler C4.5 algoritması temeline dayanan bir karar ağacı algoritmasıdır. Karar ağaçları bir makine öğrenmesi algoritmasından bilgi temsil etmede klasik bir yoldur ve veri yapılarını ifade etmekte güçlü ve hızlı bir yol sunar. Bu algoritma verileri özyinelemeli olarak sınıflandırır. Bu işlem eğitim verilerinin maksimum doğruluğunu sağlar ama verilerin sadece belirli davranış özelliklerini tanımlayan aşırı kurallar oluşturabilir [19].

3.2. Kullanılan Metrikler

3.2.1. Doğruluk–Hata Oranı (Accuracy-Error Rate)

Model başarımının ölçülmesinde kullanılan en popüler ve basit yöntem, modele ait doğruluk oranıdır. Doğru sınıflandırılmış örnek sayısının (TP +TN), toplam örnek sayısına (TP+TN+FP+FN) oranıdır. Hata oranı ise bu değerin 1'e tamlayanıdır. Diğer bir ifadeyle yanlış sınıflandırılmış örnek sayısının (FP+FN), toplam örnek sayısına (TP+TN+FP+FN) oranıdır [21].

$$\text{Doğruluk} = \frac{(TP+TN)}{(TP+FP+FN+TN)}$$

$$\text{Hata Oranı} = \frac{(FP+FN)}{(TP+FP+FN+TN)}$$

3.2.2. Kesinlik (Precision)

Kesinlik, sınıfı 1 olarak tahmin edilmiş True Pozitif (TP) örnek sayısının, sınıfı 1 olarak tahmin edilmiş tüm örnek sayısına (TP+FP) oranıdır [21].

$$\text{Kesinlik} = \frac{TP}{(TP+FP)}$$

3.2.3. Duyarlılık (Recall)

Doğru sınıflandırılmış pozitif örnek (TP) sayısının, toplam pozitif örnek sayısına (TP+FN) oranıdır [21].

$$\text{Duyarlılık} = \frac{TP}{(TP+FN)}$$

3.2.4. F-Ölçütü (F-Measure)

Kesinlik ve duyarlılık ölçütleri tek başına anlamlı bir karşılaştırma sonucu çıkarmamıza yeterli değildir. Her iki ölçütü beraber değerlendirmek daha doğru sonuçlar verir. Bunun için f-ölçütü (F) tanımlanmıştır. F-ölçütü, kesinlik (K) ve duyarlılığın (D) harmonik ortalamasıdır [21].

$$F = \frac{2DK}{(D+K)}$$

3.2.5. Kappa İstatistiği

Gözlemciler arası varyasyon, iki veya daha fazla bağımsız gözlemciler tarafından aynı şeyi değerlendiriyor olduğu her durumda ölçülebilir [9]. Kappa katsayısı -1 ile +1 arasında değişir. Tam uyum söz konusu olduğunda K=1 olur. Gözlenen uyumun şansa bağlı uyuma eşit ya da ondan büyük olması durumunda K≥0 iken, gözlenen uyumun şansa bağlı uyumundan küçük olması durumunda K<0 olur. Kappa katsayısının yorumlanabilir aralığı 0 ile +1 arasında olup, negatif (K<0) değerlerinin güvenilirlik açısından bir anlamı yoktur. 0.4 üzerindeki bir kappa skoru makul bir anlaşmayı ifade eder [22]. Kappa değeri şu şekilde hesaplanır [21].

$$K = \frac{(P_o - P_c)}{(1 - P_c)}$$

(Po kabul edilen oran, Pc kabul edilmesi beklenen oran)

4. Sonuçlar

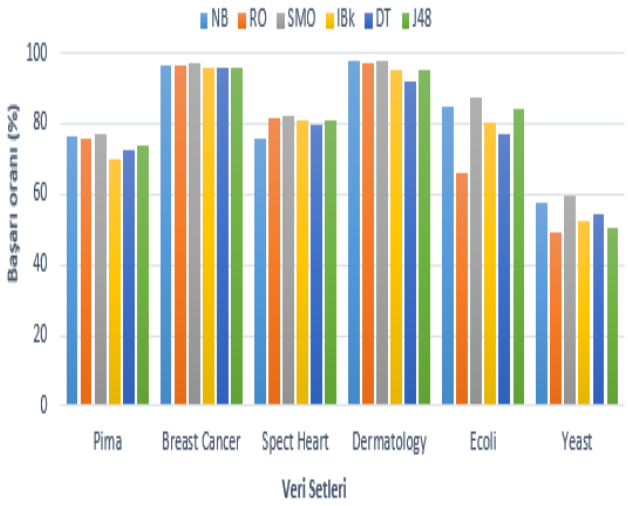
UCI veri kümesi üzerinden elde edilen 6 farklı biyomedikal veriler ile WEKA programı üzerinde makine öğrenmesi algoritmaları çalıştırılmıştır. Gerçekleştirilen 10k çapraz geçerleme sonucunda Çizelge 2' deki sonuçlara ulaşılmıştır.

	Doğruluk-Accuracy (%)					
	NB	RO	SMO	IBk	DT	J48
Pima	76.3	75.78	**77.34**	70.18	72.39	73.82
Breast Cancer	96.33	96.77	**97.07**	95.75	95.6	96.04
Spect Heart	75.65	81,27	**82.39**	80.89	79.4	80.89
Dermatology	97.48	97.2	**97.76**	95.25	91.89	95.25
Ecoli	85.11	65.77	**87.2**	80.35	76.78	84.22
Yeast	57.61	49.12	**59.7**	52.29	54.51	50.33

Çizelge 2 Çalışma Sonuçları –Doğruluk Metriği

	Hata-Error (%)					
	NB	RO	SMO	IBk	DT	J48
Pima	23.7	24.22	**22.66**	29.82	27.61	26.18
Breast Cancer	3.67	3.23	**2.93**	4.25	4.4	3.96
Spect Heart	24.35	18.73	**17.61**	19.11	20.6	19.11
Dermatology	2.52	2.8	**2.24**	4.75	8.11	4.75
Ecoli	14.89	34,23	**12.8**	19.65	23.22	15.78
Yeast	42.39	50.88	**40.3**	47.71	45.49	49.67

Çizelge 3 Çalışma Sonuçları –Hata Metriği



Şekil 3 Çalışma Sonuçlarının Grafiği

Dermatology veri seti için Doğruluk dışındaki metriklerin değerleri Çizelge 4'te gösterilmektedir.

	Dermatology Veri Seti için			
	P	R	F	Kappa
NB	0.977	0.975	**0.975**	0.9685
RO	0.972	0.972	**0.972**	0.9649
SMO	0.978	0.978	**0.978**	0.972
IBk	0.955	0.953	**0.953**	0.9405
DT	0.922	0.919	**0.919**	0.898
J48	0.953	0.953	**0.953**	0.9405

Çizelge 4 Çalışma Sonuçları – P-R-F-Kappa Metirkleri

5. Tartışma

Yapılan çalışma sonucunda Biyomedikal veriler üzerinde Makine Öğrenmesi yöntemleri test edilmiştir. Gerçekleştirilen deneyde çapraz doğrulama ile sistemin eğitim ve test performansı ölçülmüştür. Çapraz geçerleme için k=10 olarak alınmıştır.

Gerçekleştirilen çalışmada SMO algoritması diğer makine öğrenmesi algoritmalardan daha yüksek bir doğruluk ile sınıflandırmayı gerçekleştirmiştir.

Özellik sayısı, Sınıf sayısı ve örnek sayısında ki değişimler SMO'nun performansını diğer makine öğrenmesi yöntemlerine göre etkilememiştir. SMO

tüm bu bileşenlerdeki değişime rağmen en yüksek başarı oranını göstermiştir.

Bu çalışmada yeni bir yöntem önerilmemiş, gerçek bir veri seti üzerindeki makine öğrenmesi yöntemlerinin performansı ölçülmüştür. Gerçekleştirilen çalışma, bir deneyimin aktarılmasıdır. Gerçekleştirilen çalışma sonucunda yeni bir yöntem önerisi getirilmemiş olsa da bundan sonraki benzer çalışmalara bir temel oluşturmuş ve üzerinde çalışılan makine öğrenmesi algoritması için karşılaştırma yapabilmemizi sağlayacak veriler sunmaktadır.

Gerçekleştirilen çalışma sonuncunda SMO algoritmasının yüksek başarısı, bundan sonra ki çalışmalarda bizim temel noktamızı oluşturacaktır.

6.Kaynaklar

[1] Schapire, R. E. , 2003. The boosting approach to machine learning: An overview. In Nonlinear estimation and classification (pp. 149-171). Springer, New York.

[2] Ayodele, T. O. , 2010. Machine learning overview, Intech Open Access Publisher.

[3] Dietterich, T. G. , 1997. Machine-learning research, AI magazine, 18(4), 97.

[4] Chao, W. L. , 2011. Machine Learning Tutorial.

[5] Gentleman, R. , Huber, W. , Carey, V. J. , 2008. Supervised machine learning - In Bioconductor Case Studies (pp. 121-136), Springer, New York.

[6] Kotsiantis, S. B. , Zaharakis, I. , Pintelas, P. , 2007. Supervised machine learning: A review of classification techniques.

[7] Afrin, F. , Nahar, I. , 2015. Incremental learning based intelligent job search system, Doktora Tezi- BRAC Üniversitesi.

[8] Caruana, R. , Niculescu-Mizil, A. , 2006. An empirical comparison of supervised learning algorithms. In Proceedings of the 23rd international conference on Machine learning (pp. 161-168), ACM.

[9] Sebastiani, F. , 2002. Machine learning in automated text categorization, ACM computing surveys (CSUR), 34(1), 1-47.

[10] Özgür, A. , 2004. Supervised and unsupervised machine learning techniques for text document categorization, Doktora Tezi-Bogaziçi Üniversitesi.

[11] Çoban, Ö. , 2016. Metin Sınıflandırma Teknikleriyle Türkçe Twitter Duygu Analizi, Yüksek Lisans Tezi, Atatürk Üniversitesi.

[12] http://bilgisayarkavramlari.sadievrenseker.com/2009/06/01/weka/ , Erişim Tarihi: 07. 07.2016.

[13] https://tr.wikipedia.org/wiki/Weka , Erişim Tarihi: 07. 07.2016.

[14] UCI, http://archive.ics.uci.edu/ml/, , Erişim Tarihi: 07. 07.2016.

[15] Karakoyun, M. , Hacıbeyoğlu, M. , 2005. Biyomedikal Veri Kümeleri ile Makine Öğrenmesi Sınıflandırma Algoritmalarının İstatistiksel olarak Karşılaştırılması, Dokuz Eylül Üniversitesi Mühendislik Fakültesi Dergisi, 16(48), (pp. 30-41).

[16] Breiman, L. , Cutler, A. , 2005. Random Forests.

[17] Breiman, L. , 2001. Machine Learning, 45, (pp. 5–32), Random Forests.

[18] Platt, J.C. , 1998. , Sequential Minimal Optimization: A Fast Algorithm for Training Support Vector Machines.

[19] Gandhi, M.G. , Srivatsa, S.K. , 2010. Classification Algorithms in Comparing Classifier Categories to Predict the Accuracy of the Network Intrusion Detection – A Machine Learning Approach. Advances in Computational Sciences and Technology, 3(3).

[20] Sehgal, L. , Mohan, N. , Sandhu, P.S. , 2012. Prediction of Function Based Software Using Decision Tree Approach.

[21] Nizam, H., Akın,S.S., 2014. Sosyal Medyada Makine Öğrenmesi ile Duygu Analizinde Dengeli ve Dengesiz Veri Setlerinin Performanslarının Karşılaştırılması, XIX. Türkiye'de İnternet Konferansı, İzmir.

[22] Landis, J. R. , Gary G.K., 1977. The measurement of observer agreement for categorical data, Biometrics 33(1), (pp 159-174).