**Özet:** Makine öğrenmesi algoritmaları veri sınıflandırılması için kullanılan yöntemlerdir. Bu çalışmanın amacı tek bir veri seti üzerinde makine öğrenmesi algoritmalarının karşılaştırılmasıdır. Bu çalışma için 3 farklı makine öğrenmesi algoritması WEKA yazılımında tek veri seti üzerinde 10 katlı çapraz doğrulama tekniği ile sınıflandırılmıştır. Sınıflandırma sonucuna göre, veri setine uygulanan makine öğrenmesi algoritmalarının göreceli mutlak hata değerleri, her makine öğrenmesi algoritması için farklı çıkmıştır. Iris veri seti için en düşük makine öğrenmesi algoritması Naive Bayes algoritması elde edilmiştir.

**Anahtar Kelime:** Makine öğrenmesi, sınıflandırma, WEKA, veri seti.

**Abstract:** Machine learning algorithms are methods used for data classification. The aim of this study is a comparison of machine learning algorithms on same dataset. For this study, 3 different machine learning algorithms with tenfold cross validation method in WEKA classified same dataset. According to the classification result, the relative absolute error values of the machine learning algorithms applied to the data set are different for each machine learning algorithm. The lowest machine learning algorithm for the Iris data set is the Naive Bayes Algorithm.

**Keywords:** Machine learning,classification, WEKA, dataset.

**1. Giriş**

Bu çalışmada makine öğrenmesi algoritmalarının tek bir veri seti üzerinde etkisi incelenmiştir.Makine öğrenmesi algoritmaları olarak Random Forest, BayesNET ve Naive Bayes ve kullanılmıştır. Ayrıca bu makine öğrenmesi algoritmaları WEKA yazılımında[7] 10 katlı çapraz doğrulama tekniğine göre sınıflandırılmıştır.Veri seti olarak ise, UCI Machine Learning Repository websitesinden alınan Iris veri seti kullanılmıştır [1].

Makine öğrenmesi algoritmalarının veri seti üzerindeki sınıflandırılmasıyla elde edilen sonuçlara göre; Random Forest algoritmasından elde edilen göreceli mutlak hata değeri %9.19, BayesNET algoritmasından elde edilen göreceli mutlak hata değeri %10.2111, ve Naive Bayes algoritmasından elde edilen göreceli mutlak hata değeri %7.6997 olarak elde edilmiştir.

Bildirinin devamında, bu konuya benzer çalışmaların yer aldığı ilgili çalışma bölümü ; WEKA yazılımı, bildiride kullanılan iris veri seti, algoritmalar ve metriklerin tanıtıldığı yöntemler bölümü; çalışmada elde edilen sonuçların yer aldığı Deneysel Sonuçlar bölümü ve son olarak da Sonuç bölümü bulunmaktadır.

**2. İlgili Çalışmalar**

Kharche[4] makine öğrenmesi yöntemlerini 5 farklı data üzerinde 4 farklı algoritma ile denemiş ve en düşük göreceli mutlak hata değerine sahip algoritma Naive Bayes algoritması olduğunu doğrulamıştır.

Ozen[5] Yapay Sinir Ağı Algoritması Kullanarak tek sınıf üzerinde 2 farklı algoritma ile karşılaştırma yapmış Naive Bayes algoritmasının göreceli mutlak hata belirlemede en iyi çözüm olacağını belirtmeştir.

Kaynar[6] Naive Bayes, veri setleri üzerinde kullanarak sonuçları karşılaştırmışlardır. Eğitim ve test verileri için Yapay Sinir Ağları ve Destek Vektör makineleri yöntemlerinde daha iyi sonuç elde etmişlerdir.

**3. Yöntemler**

**3.1. WEKA**

WEKA, veri madenciliği çalışmaları için kullanılan makine öğrenmesi algoritmalarını içeren açık kodlu bir yazılımdır [2] Bu çalışmada, veri setleri WEKA yazılımında farklı makine öğrenmesi algoritmalarıyla sınıflandırılmıştır.

**3.2. Veri Seti**

Çalışma için tek veri seti kullanılmıştır.Bu veri seti Iris veri setidir [3].

Iris veri seti: Bu veri seti çiçek türlerinin çanak yaprağı ve taç yapraklarının genişlik ve uzunluk özelliklerini kapsar. Bu veri setinde 150 örnek ve girdi vardır.

**3.2. Algoritmalar ve Metrikler**

Bu çalışmada kullanılan algoritmalar aşağıda verilmiştir. Algoritmaları kullanırken varsayılan algoritma parametreleri kullanılmış değişiklik yapılmamıştır.

• RandomForest -P 100 -I 100 -num-slots 1 -K 0 -M 1.0 -V 0.001 -S 1

Kullanılan parametrelerin anlamları aşağıda açıklanmaktadır.

-P:yüzde olarak eğitim setirin boyutunu ayarlar.(varsayılan: 100)

-I <num>:Yineleme sayısı.(varsayılan: 100)

-num-slots <say>:  paralel yürütme yuvalarının sayısını ifade etmektedir.(varsayılan: 1)

-K <say>: Rastgele araştırılacak niteliklerin sayısı.(varsayılan: 0)

-M <say>: Yaprak başına asgari örnek sayısını ayarlayın. (varsayılan: 1)

-V <minimum varyans>  Minimum sayısal sınıf varyans oranını ayarlar (varsayılan 1e-3).

-S <say>: Rasgele sayı üreteci için tohum sayısı.(varsayılan: 1) [8]

• BayesNet -D -Q weka.classifiers.bayes.net.search.local.K2 -- -P 1 -S BAYES -E weka.classifiers.bayes.net.estimate.SimpleEstimator -- -A 0.5

Kullanılan parametrelerin anlamları aşağıda açıklanmaktadır.

-D:ADTree veri yapısını kullanma

  -Q weka. classifiers. bayes. net. search. SearchAlgorithm: Arama algoritması

  -E weka. classifiers. bayes. net. estimate.SimpleEstimator: Tahmin algoritması

• NaiveBayes

Bu çalışmada kullanılan parametreler aşağıda verilmiştir.

• Sepallength

• Sepalwidth

• Petallength

• Petalwidth

• class

Bu çalışmada kullanılan metrik aşağıda verilmiştir.

Göreceli Mutlak Hata Değeri (Relative Absolute Error) : Gerçek değere ne kadar yaklaşıldığının oransal gösteren bir hata çeşididir.Hata, tahmin edilen değer ile gerçek değer arasındaki farktır.

**4. Araştırma Bulguları**

Bu çalışmada Random Forest, BayesNET ve Naive Bayes algoritmalarının veri seti üzerindeki etkisi incelenmiştir. Bu çalışmaya göre elde edilen göreceli mutlak hata değerleri Şekil 1’de gösterilmiştir.

A picture containing screenshot

Description generated with very high confidence

Şekil 1. Göreceli Mutlak Hata Değerleri

Iris veri seti için en yüksek göreceli mutlak hata değeri %10.2111 ile BayesNET algoritması, en düşük göreceli mutlak hata değeri %7.6997 ile Naive Bayes algoritmasıdır. Tablo 3’de bu çalışmada kullanılan tüm algoritmaların göreceli mutlak hata değerleri verilmiştir.

**Tablo 1.** Göreceli Mutlak Hata Değerleri Tablosu

|  |  |
| --- | --- |
| **Algoritmalar** | **Göreceli Mutlak Hata Değerleri(%)** |
| Random Forest | 9.19 |
| BayesNET | 10.2111 |
| Naive Bayes | 7.6997 |

Algoritmalar Göreceli Mutlak Hata Değerleri(%)

Şekil 2’de veri setine göre göreceli mutlak hata değerleri verilmiştir. Buna göre Iris veri setinde en düşük göreceli mutlak hata değerine sahip makine öğrenmesi algoritması Naive Bayes algoritması ve değeri %7.6997’dir.

A screenshot of a cell phone

Description generated with very high confidence**Şekil 2.** Göreceli mutlak hata değerleri

**5. Sonuç**

Makine öğrenmesi algoritmalarının aynı veri seti üzerinde karşılaştırılmasıyla farklı göreceli mutlak hata değerleri elde edilmiştir. Bu çalışma için kullanılan algoritmalar; Random Forest, BayesNET ve Naive Bayes algoritmalarıdır. Kullanılan veri seti ise UCI Machine Learning Data Repository websitesinden alınan Iris veri setidir.

Sınıflandırma sonucuna göre, en düşük göreceli mutlak hata değerine sahip algoritma Naive Bayes algoritmasıdır ve değeri %7.6997 olarak saptanmıştır.

**REFERENCES**

[1] Lichman, M. (2013). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science, Alınma Tarihi: 09/01/2017.

[2] Weka 3: Data Mining Software İn Java, Web Sayfası: Https://Www.Cs.Waikato.Ac.Nz/Ml/Weka/, Alınma Tarihi: 09/01/2017.

[3] Iris Data Set, Web Sayfası: Http://Archive.İcs.Uci.Edu/Ml/Datasets/Iris, Alınma Tarihi: 09/01/2017.

[4] Kharche D, Comparıson Of Dıfferent Datasets Usıng Varıous Classıfıcatıon Technıques Wıth Weka, Inc., Vol. 3, Issue. 4, 389 – 393, 2014

[5] Ozen Z, R İle Geri Yayılım Yapay Sinir Ağı Algoritması Kullanarak Bir Sınıflandırma Uygulaması, R İle Veri Madenciliği Uygulamaları Bölüm 4, 2016

[6] Kaynar O, Sentiment Analysis With Machine Learning Techniques, International Artificial Intelligence And Data Processing Symposium, 2016

[7] Eibe Frank, Mark A. Hall, and Ian H. Witten (2016). The WEKA Workbench. Online Appendix for "Data Mining: Practical Machine Learning Tools and Techniques", Morgan Kaufmann, Fourth Edition, 2016.

[8] George, H.J. and Pat, L., Estimating Continuous Distributions in Bayesian Classifiers, 1995, Web sayfası: <http://weka.sourceforge.net/doc.dev/weka/classifiers/trees/RandomForest.html>, Alınma Tarihi: 27/01/2017.

[9] George, H.J. and Pat, L., Estimating Continuous Distributions in Bayesian Classifiers, 1995, Web sayfası: http://weka.sourceforge.net/doc.dev/weka/classifiers/bayes/NaiveBayes.html, Alınma Tarihi: 27/01/2017.