Makine Öğrenmesi Algoritmalarının Farklı Veri Setleri Üzerinde Karşılaştırılması
Comparison of Machine Learning Algorithms on Different Datasets

Elif UYSAL
Bilgisayar Mühendisliği
KTO Karatay Üniversitesi
Konya, Türkiye
elif.uysal@ogrenci.karatay.edu.tr

Ali ÖZTÜRK
Bilgisayar Mühendisliği
KTO Karatay Üniversitesi, Konya, Türkiye
Havelsan A.Ş., Ankara, Türkiye
ali.ozturk@karatay.edu.tr

***Özetçe* —Makine öğrenmesi algoritmaları verileri sınıflandırmak için kullanılan yöntemlerdir. Bu çalışmanın amacı, farklı veri setleri üzerinde makine öğrenmesi algoritmalarının karşılaştırılmasıdır. Bu çalışma için 9 farklı makine öğrenmesi algoritması WEKA yazılımında 10 katlı çapraz doğrulama tekniği ile 3 farklı veri seti üzerinde sınıflandırılmıştır. Sınıflandırma sonucuna göre, yüksek doğruluk değerine sahip makine öğrenmesi algoritması 3 veri seti için farklı çıkmıştır. Araba Değerlendirme (Car Evaluation) veri seti için Multilayer Perceptron algoritması, Resim Segmantasyonu (Image Segmentation) veri seti için Random Forest algoritması ve Kullanıcı Bilgi Modelleme (User Knowledge Modeling) veri seti için Simple Logistic algoritması elde edilmiştir.**

Anahtar Kelimeler—veri seti, makine öğrenmesi, sınıflandırma, WEKA.

***Abstract*—Machine learning algorithms are methods used to classify data. Aim of this study is comparison of machine learning algorithms on different datasets. For this study, 9 different machine learning algorithms with 10 fold cross validation method in WEKA is classified on 3 different datasets. As a result of classification, machine learning algorithm which has high accuracy rate is different for 3 datasets. Multilayer Perceptron algorithm for Car Evaluation dataset, Random Forest algorithm for Image Segmentation dataset and Simple Logistic algorithm for User Knowledge Modeling dataset were obtained.**

Keywords—dataset, machine learning, classification, WEKA.

I. GİRİŞ

Günümüzde verilerin işlenmesi ve sınıflandırılması makine öğrenmesi yöntemleri ile kolaylık kazanmıştır. Bu çalışmada, makine öğrenmesi algoritmalarının farklı veri setleri üzerinde etkisi incelenmiştir. Makine öğrenmesi algoritmaları olarak `Multilayer Perceptron`, `Naive Bayes`, `Logit Boost`, `J48`, `Random Forest`, `Bayes Net`, `Simple Logistic`, `Bagging` ve `Logistic Model Trees (LMT)` kullanılmıştır. Ayrıca bu makine öğrenmesi algoritmaları WEKA yazılımında 10 katlı çapraz doğrulama tekniğine göre sınıflandırılmıştır. Veri seti olarak ise, UCI Machine Learning Repository[1] websitesinden alınan `Araba Değerlendirme`, `Resim Segmantasyonu` ve `Kullanıcı Bilgi Modelleme` veri setleri kullanılmıştır.

Makine öğrenmesi algoritmalarının farklı veri setleri üzerinde sınıflandırılmasıyla elde edilen sonuçlara göre; Araba Değerlendirme veri setinde doğruluk değeri **%99.537** ile *Multilayer Perceptron* algoritması, Resim Segmantasyonu veri setinde **%98.0519** ile *Random Forest* algoritması ve Kullanıcı Bilgi Modelleme veri setinde **%95.7816** ile *Simple Logistic* algoritması elde edilmiştir.

Bildirinin devamında, bu konuya benzer çalışmaların yer aldığı İlgili Çalışma bölümü; WEKA yazılımı, bildiride kullanılan veri setleri, algoritmalar ve metriklerin tanıtıldığı Yöntemler bölümü; çalışmada elde edilen sonuçların yer aldığı Deneysel Sonuçlar bölümü ve son olarak da Sonuç bölümü bulunmaktadır.

II. İLGİLİ ÇALIŞMA

Bilgin [1] makine öğrenmesi yöntemlerini 10 parça çapraz doğrulama tekniği ile 6 farklı biyomedikal veri seti üzerindeki performans analizini incelemiştir. Bilgin [1] makine öğrenmesi algoritması olarak Naive Bayes, Random Forest, Sequential Minimal Optimization (SMO), IBk (k-nearest neighbor), Decision Table, J48 kullanmıştır ve SMO algoritması ile daha yüksek doğruluk değeri elde etmiştir.

Akçetin vd. [2] ADTree, BFTree, Fonksiyonel Ağaçlar, J48, J48 Aşı, LADTree, LMT, NBTree, Rastgele Orman Karar Ağacı, Rastgele Ağaç, REPTree Sınıflandırıcısı, SimpleCART algoritmaları 10 katlı çapraz doğrulama ile UCI elektronik posta veri seti üzerinde sınıflandırmıştır ve Rastgele Orman karar ağacı %94.68 ile en iyi sınıflandırma başarısı elde etmiştir.

Caruana vd. [3] SVMs, neural nets, logistic regression, naivebayes, memory based learning, random forests, decision trees, bagged trees, boosted trees ve boosted stumps algoritmalarını 11 tane veri seti üzerinde sınıflandırmışlardır. Ayrıca

978-1-5386-1501-0/18/$31.00 © 2018 IEEE

[1] https://archive.ics.uci.edu/ml/datasets.html

Platt Scaling ve Isotonic Regression yöntemlerini de kullanarak bir karşılaştırma yapmışlardır. Bu iki yöntemin kullanılmasıyla en iyi algoritma boosted trees algoritması elde edilmiştir, normal sınıflandırmanın yapılmasıyla en iyi algoritma bagged trees algoritması elde edilmiştir.

Kaynar vd. [4] Naive Bayes, Merkez Tabanlı Sınıflayıcı, Çok Katmanlı Yapay Sinir Ağları ve Destek Vektör Makineleri yöntemlerini film yorumları veri seti üzerinde kullanarak sonuçları karşılaştırmışlardır. Eğitim ve test verileri için Yapay Sinir Ağları ve Destek Vektör makineleri yöntemlerinde daha iyi sonuç elde etmişlerdir.

III. YÖNTEMLER

A. WEKA

WEKA [9], veri madenciliği çalışmaları için kullanılan, makine öğrenmesi algoritmalarını içeren açık kaynak kodlu bir yazılımdır.[2] Bu çalışmada, veri setleri WEKA yazılımında farklı makine öğrenmesi algoritmalarıyla sınıflandırılmıştır.

B. Veri Seti

Çalışma için üç farklı veri seti kullanıldı. Bunlar Tablo I'de de görüldüğü gibi Araba Değerlendirme [7], Resim Segmantasyonu [7] ve Kullanıcı Bilgi Modelleme [8] veri setleridir.

Tablo I: Veri seti detayları

veri seti	özellik sayısı	örnek sayısı
Araba Değerlendirme	7	1728
Resim Segmantasyonu	20	2310
Kullanıcı Bilgi Modelleme	6	403

Araba Değerlendirme veri seti[3]: Bu veri seti araba ile ilgili özellikleri kapsar. Örneğin, arabanın kapı sayısı, kişi kapasitesi, güvenirliliği gibi özellikleri içerir. Bu veri setinde 1728 örnek vardır.

Resim Segmantasyonu veri seti[4]: Bu veri seti 7 dış mekan resimlerinden rastgele çekilen örneklerden oluşur. Bu veri setinde 20 özellik ve 2310 örnek vardır.

Kullanıcı Bilgi Modelleme veri seti[5]: Yazarların sezgisel bilgi sınıflandırıcı, k-en yakın komşu algoritmasını kullanarak kullanıcıların bilgi sınıfını sınıflandırmasıyla ilgili veri setlerini kapsar. Bu veri seti 6 özellik ve 403 örnekten oluşur.

C. Algoritmalar ve Metrikler

Multilayer Perceptron, Naive Bayes, Logit Boost, J48, Random Forest, Bayes Net, Simple Logistic, Bagging ve LMT algoritmaları WEKA yazılımında varsayılan özellikler kullanılarak sınıflandırıldı.

- **Multilayer Perceptron:** Öğrenme oranı (learning rate) *0.3* ve gizli katmanda *5* düğüm kullanıldı.
- **Naive Bayes:** Denetimli ayrıştırma (supervised discretization) *false* kullanıldı.
- **Logit Boost:** Sınıflandırıcı olarak *Decision Stump* algoritması kullanıldı.
- **J48:** Güven faktörü (confidence factor) *0.25* kullanıldı.
- **Random Forest:** Bag size percent değeri olarak 100 kullanıldı.
- **Bayes Net:** Tahmin (estimator) algoritması olarak *Simple Estimator* algoritması ve arama (search) algoritması olarak *K2* algoritması kullanıldı.
- **Simple Logistic:** Boosting iteration sayısı *0* olarak kullanıldı.
- **Bagging:** Sınıflandırıcı olarak *REPTree* algoritması kullanıldı.
- **LMT:** Boosting iteration sayısı *-1* olarak kullanıldı.

Bu çalışmada kullanılan metrikler aşağıda verilmiştir.

Tablo II'ye göre [5], TP (true positive) ve FN (false negative) değerleri doğru ve yanlış örneklerin doğru şekilde sınıflandırıldığını gösterir; FP (false positive) ve TN (true negative) değerleri ise doğru ve yanlış örneklerin doğru şekilde sınıflandırılmadığını gösterir.

Tablo II: Karışıklık Matrisi [5]

	Beklenen Pozitif	Beklenen Negatif
Gerçek Doğru	TP	FN
Gerçek Yanlış	FP	TN

- **Doğruluk (accuracy) [4]:** Doğru olarak sınıflandırılan örneklerin toplam örnek sayısına oranıdır.

$$doğruluk = \frac{TP + TN}{TP + TN + FP + FN} \quad (1)$$

- **Kesinlik (precision) [4]:** Pozitif olarak etiketlenen örneklerin sayısının pozitif olarak sınıflandırılan toplam örneklere oranıdır.

$$kesinlik = \frac{TP}{TP + FP} \quad (2)$$

- **Duyarlılık (recall) [4]:** Pozitif olarak etiketlenen örneklerin gerçekten pozitif olan örneklerin toplam sayısına eşittir.

$$duyarlılık = \frac{TP}{TP + FN} \quad (3)$$

- **F-ölçütü (f-measure) [4]:** Kesinlik ve duyarlılık değerlerinin harmonik ortalamasıdır.

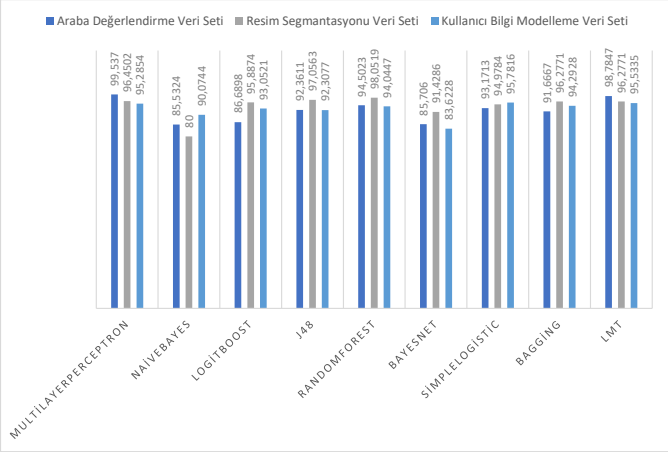
$$f - ölçütü = \frac{2 * kesinlik * duyarlılık}{kesinlik + duyarlılık} \quad (4)$$

- **Ortalama mutlak hata (mean absolute error) [6]:** Mutlak hataların ortalamasıdır. Hata, tahmin edilen değer ile gerçek değer arasındaki farktır.

[2]https://www.cs.waikato.ac.nz/ml/weka/
[3]https://archive.ics.uci.edu/ml/datasets/Car+Evaluation
[4]https://archive.ics.uci.edu/ml/datasets/Image+Segmentation
[5]https://archive.ics.uci.edu/ml/datasets/User+Knowledge+Modeling

IV. Deneysel Sonuçlar

Bu çalışmada, Multilayer Perceptron, Naive Bayes, Logit Boost, J48, Random Forest, Bayes Net, Simple Logistic, Bagging ve LMT algoritmalarının farklı veri setleri üzerinde etkisi incelenmiştir. Bu çalışmaya göre elde edilen doğruluk değerleri Şekil 1'de gösterilmiştir.



Şekil 1: Doğruluk değerleri

Araba Değerlendirme veri seti için en yüksek doğruluk değeri **%99.537** ile `Multilayer Perceptron` algoritması, en düşük doğruluk değeri **%85.5324** ile `Naive Bayes` algoritmasıdır. Resim Segmantasyonu veri seti için en yüksek doğruluk değeri **%98.0519** ile `Random Forest` algoritması, en düşük doğruluk değeri **%80** ile `Naive Bayes` algoritmasıdır. Kullanıcı Bilgi Modelleme veri seti için en yüksek doğruluk değeri **%95.7816** ile `Simple Logistic` algoritması, en düşük doğruluk değeri **%83.6228** ile `Bayes Net` algoritmasıdır. Sonuç olarak, doğruluk değerlerinin yüksek olduğu algoritmalar üç veri seti için de farklıdır; fakat, doğruluk değerlerinin düşük olduğu algoritmalar Araba Değerlendirme veri seti ve Resim Segmantasyonu veri seti için aynıdır.

Tablo III'te Araba Değerlendirme veri seti için elde edilen ağırlıklı ortalama kesinlik, ağırlıklı ortalama duyarlılık ve ağırlıklı ortalama f-ölçütü değerleri verilmiştir. Bu tabloya göre, `Multilayer Perceptron` algoritması **0.995** değeri ile en yüksek kesinlik, duyarlılık ve f-ölçütü değerlerine sahiptir. `Naive Bayes` algoritması ise **0.852**, **0.855** ve **0.847** değerleri ile sırasıyla en düşük kesinlik, duyarlılık ve f-ölçütü değerlerine sahiptir.

Tablo III: Araba Değerlendirme veri seti için sonuçlar

Algoritmalar	Kesinlik	Duyarlılık	F-Ölçütü
MultilayerPerceptron	**0.995**	**0.995**	**0.995**
NaiveBayes	0.852	0.855	0.847
LogitBoost	0.869	0.867	0.861
J48	0.924	0.924	0.924
RandomForest	0.946	0.945	0.945
BayesNet	0.854	0.857	0.849
SimpleLogistic	0.933	0.932	0.932
Bagging	0.919	0.917	0.916
LMT	0.988	0.988	0.988

Tablo IV'te Resim Segmantasyonu veri seti için elde edilen ağırlıklı ortalama kesinlik, ağırlıklı ortalama duyarlılık ve ağırlıklı ortalama f-ölçütü değerleri verilmiştir. Bu tabloya göre, `Random Forest` algoritması **0.981**, **0.981** ve **0.980** değerleri ile sırasıyla en yüksek kesinlik, duyarlılık ve f-ölçütü değerlerine sahiptir. `Naive Bayes` algoritması ise **0.817**, **0.800** ve **0.779** değerleri ile sırasıyla en düşük kesinlik, duyarlılık ve f-ölçütü değerlerine sahiptir.

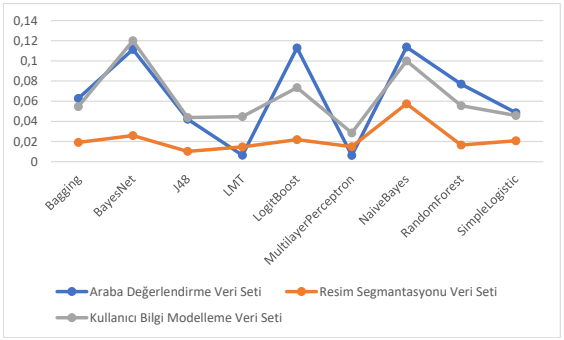
Tablo IV: Resim Segmantasyonu veri seti için sonuçlar

Algoritmalar	Kesinlik	Duyarlılık	F-Ölçütü
MultilayerPerceptron	0.965	0.965	0.964
NaiveBayes	0.817	0.800	0.779
LogitBoost	0.959	0.959	0.959
J48	0.971	0.971	0.971
RandomForest	**0.981**	**0.981**	**0.980**
BayesNet	0.914	0.914	0.914
SimpleLogistic	0.951	0.950	0.950
Bagging	0.963	0.963	0.963
LMT	0.963	0.963	0.963

Tablo V'te Kullanıcı Bilgi Modelleme veri seti için elde edilen ağırlıklı ortalama kesinlik, ağırlıklı ortalama duyarlılık ve ağırlıklı ortalama f-ölçütü değerleri verilmiştir. Elde edilen bu sonuçlara göre, `Simple Logistic` algoritması **0.958** değeri ile en yüksek kesinlik, duyarlılık ve f-ölçütü değerine sahiptir. `Bayes Net` algoritması ise, **0.842**, **0.836** ve **0.837** değerleri ile sırasıyla en düşük kesinlik, duyarlılık ve f-ölçütü değerlerine sahiptir.

Tablo V: Kullanıcı Bilgi Modelleme veri seti için sonuçlar

Algoritmalar	Kesinlik	Duyarlılık	F-Ölçütü
MultilayerPerceptron	0.953	0.953	0.953
NaiveBayes	0.908	0.901	0.901
LogitBoost	0.931	0.931	0.931
J48	0.923	0.923	0.923
RandomForest	0.942	0.940	0.940
BayesNet	0.842	0.836	0.837
SimpleLogistic	**0.958**	**0.958**	**0.958**
Bagging	0.945	0.943	0.943
LMT	0.956	0.955	0.955



Şekil 2: Ortalama mutlak hata değerleri

Şekil 2'de veri setlerine göre ortalama mutlak hata değerleri verilmiştir. Buna göre, Araba Değerlendirme veri setinde `Multilayer Perceptron` algoritması en düşük değere sahip ve değeri **0.0062**, Resim Segmantasyonu veri

setinde `J48` algoritması en düşük değere sahip ve değeri **0.0102**, Kullanıcı Bilgi Modelleme veri setinde `Multilayer Perceptron` algoritması en düşük değere sahip ve değeri **0.0286**'dır.

V. SONUÇ

Makine öğrenmesi algoritmalarının farklı veri setleri üzerinde karşılaştırılmasıyla farklı sonuçlar elde edilmiştir. Bu çalışma için kullanılan algoritmalar; Multilayer Perceptron, Naive Bayes, Logit Boost, J48, Random Forest, Bayes Net, Simple Logistic, Bagging ve Logistic Model Trees (LMT) algoritmalarıdır. Kullanılan veri setleri ise UCI Machine Learning Data Repository websitesinden alınan Araba Değerlendirme, Resim Segmantasyonu ve Kullanıcı Bilgi Modelleme veri setleridir.

Sınıflandırma sonucuna göre, Multilayer Perceptron algoritması Araba Değerlendirme veri setinde %99.537 doğruluk değeri ile, Random Forest algoritması Resim Segmantasyonu veri setinde %98.0519 doğruluk değeri ile, Simple Logistic algoritması Kullanıcı Bilgi Modelleme veri setinde %95.7816 doğruluk değeri ile en yüksek değere sahip algoritmalardır. Ayrıca, Araba Değerlendirme veri seti RBFNetwork ve LibSVM algoritmalarıyla sınıflandırıldığında doğruluk değerleri sırasıyla %88.2523 ve %91.8981, Resim Segmantasyonu veri seti RBFNetwork ve LibSVM algoritmalarıyla sınıflandırıldığında doğruluk değerleri sırasıyla %87.2727 ve %63.8528, Kullanıcı Bilgi Modelleme veri seti RBFNetwork ve LibSVM algoritmalarıyla sınıflandırıldığında doğruluk değerleri sırasıyla %94.5409 ve %81.6377 elde edilmiştir. Sonuç olarak, kullanılan 3 veri seti RBFNetwork ve LibSVM algoritmalarıyla sınıflandırıldığında doğruluk değerleri Multilayer Perceptron, Random Forest ve Simple Logistic algoritmalarından daha düşük elde edilmiştir.

KAYNAKÇA

[1] Bİlgİn, M., "Gerçek Veri Setlerinde Klasik Makine Öğrenmesi Yöntemlerinin Performans Analizi," Breast, vol. 2, no. 9, pp. 683–688, Şubat 2017.

[2] Akçetin, E. and Çelik, U., "The Performance Benchmark of Decision Tree Algorithms for Spam e-mail Detection," J. Internet Appl. Manag., vol. 5, no. 2, pp. 43–56, 2014.

[3] Caruana, R. and Niculescu-Mizil, A., "An empirical comparison of supervised learning algorithms," Proc. 23rd Int. Conf. Mach. Learn., vol. C, no. 1, pp. 161–168, 2006.

[4] Kaynar, O., Görmez, Y., Yıldız, M. and Albayrak, A., "Makine Öğrenmesi Yöntemleri ile Duygu Analizi Sentiment Analysis with Machine Learning Techniques," International Artificial Intelligence and Data Processing Symposium (IDAP'16), pp. 234-241, September, 2016.

[5] https://docs.microsoft.com/en-us/azure/machine-learning/studio/evaluate-model-performance

[6] https://docs.microsoft.com/tr-tr/azure/machine-learning/studio/create-experiment

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

[8] Kahraman, H. T., Sagiroglu, S., Colak, I., Developing intuitive knowledge classifier and modeling of users' domain dependent data in web, Knowledge Based Systems, vol. 37, pp. 283-295, 2013.

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