*2 Makine Öğrenmesi Algoritmasının Diyabet Veri Seti Üzerinde Karşılaştırılması*

*Comparison of 2 Machine Learning Algorithms on Diabetes Dataset*

*Burak UZMAN1\*, Ali ÖZTÜRK2*

*1Elektrik ve Bilgisayar Mühendisliği Bölümü, Fen Bilimleri Enstitüsü, Karatay Üniversitesi, Konya, Türkiye*

*2Bilgisayar Mühendisliği Bölümü, Mühendislik Fakültesi, Karatay Üniversitesi, Konya, Türkiye*

*\*Sorumlu Yazar: burakuzmaneeekto@gmail.com*

**Özet:** Bu çalışmada verileri sınıflandırmak için WEKA adlı yazılımda makine öğrenmesi algoritmalarından 2 tanesi kullanılmış ve karşılaştırılmıştır. WEKA’da diyabet datası seçilmiş ve bu data Bayes ve RandomForest algoritmalarıyla 10 katlı çapraz doğrulama tekniği ile sınıflandırılmıştır. Bu işlemden sonra ise WEKA’da select attributes ile yöntem tarafından yeterli olabilecek özellikler bulunmuş ve bu özellikler ile aynı algoritmalarla tekrar sınıflandırma yapılmıştır.

**Anahtar Kelime:** Makine öğrenmesi, sınıflandırma, WEKA, diyabet.

**Abstract:** In this study, in the software called WEKA, 2 machine learning algorithms have been used and compared for classify data. Diabetes data were selected in WEKA and this data was classified by Bayes and RandomForest algorithms using 10-fold cross-validation technique. After this operation, the WEKA has the attributes that can be satisfied by the select attributes and the method, and the same algorithms are re-classified.

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

1. Giriş

 Makine öğrenmesi esas olarak 1959 yılında [bilgisayar](http://www.endustri40.com/turkiyenin-ilk-bilgisayari-ibm-650/) biliminin yapay zekada sayısal öğrenme ve model tanıma çalışmalarından geliştirilmiş bir alt dalıdır. [Makine](http://www.endustri40.com/karanlik-fabrikalar-ile-insansiz-uretim/) öğrenmesi yapısal işlev olarak öğrenebilen ve veriler üzerinden tahmin yapabilen algoritmaların çalışma ve inşalarını araştıran bir sistemdir. Bu tür algoritmalar statik program talimatlarını harfiyen takip etmek yerine örnek girişlerden veri tabanlı tahminleri ve kararları gerçekleştirebilmek amacıyla bir model inşa ederek çalışırlar [1]. Makine öğrenmesinin birkaç öğrenme tekniği vardır. Bunlar; denetimli, denetimsiz, yarı denetimli, takviyeli ve yoğun öğrenme tekniğidir. Bu çalışmada makine öğrenmesi algoritmalarının test edilebilmesi için WEKA adlı paket programı kullanılmıştır. WEKA verileri basit bir dosyadan okur ve veriler üzerindeki [stokastik değişkenlerin](http://www.bilgisayarkavramlari.com/2009/06/17/stokastik-surec-stochastic-process/) sayısal veya nominal değerler olduğunu kabul eder. WEKA üzerinde makine öğrenmesi ve istatistik ile ilgili pekçok kütüphane hazır olarak gelmektedir [2].  Bu çalışmada WEKA programından bir data set seçilmiş ve seçilen data setine 2 farklı algoritma uygulanarak sınıflandırılmıştır.

2. Yöntemler

2.1. WEKA

Bu çalışmada WEKA programından bir veri seti seçilmiş ve bu veri setlerinde farklı algoritmalar ile sınıflandırma yapılmıştır.

2.2. Veri Seti

Çalışma için kullanılan veri seti diyabet veri setidir.

Diyabet veri seti hasta adaylarının yaşına, deri özelliklerine, hamile olma durumuna(bayansa) göre hastalığı belirleyen bir settir.

2.2. Algoritmalar ve Metrikler

* BayesNet [3]
* RandomForest [4]

Bu çalışmada kullanılan metrikler ise;

* Correctly Classified (CC)
* Mean absolute error (MAE)
* Root mean squared error (RMSE)
* Relative absolute error (RAE)
* Root relative squared error (RRSE)
* Confusion Matrix
* ROC Area

3. Benzetim Sonuçları

Bu çalımada select attributes yapılmadan önce veri setinin bütün özellikleri sınıflandırmaya dahil edilmiştir. 2 farklı algoritmanın kullanılan metriklere göre sınıflandırmaları tablo 1’de verilmiştir.

***Tablo1: Algoritma ve Metrik Karşılaştırmaları( Select Attributes Yapılmadan* ö*nce)***

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Metrik  Algoritma | CC | MAE | RMSE | RAE | RRSE | ROC |
| BayesNet | 74.349 % | 0.2987 | 0.4208 | 65.7116 % | 88.28 % | 0.806 |
| RandomForest | 75.7813 % | 0.3106 | 0.4031 | 68.3405 % | 84.5604 % | 0.820 |

Tabloya göre RandomForest algoritmasının veri setini doğru hesaplama yüzdesi BayesNet algoritmasına göre daha iyidir.

2 algoritmanın confusion matrix karşılaştırılmaları tablo 2 ve 3’te verilmiştir.

|  |  |
| --- | --- |
| a | b |
| 408 | 92 |
| 105 | 163 |

***Tablo 2: BayesNet Confusion Matrix***

|  |  |
| --- | --- |
| a | b |
| 418 | 82 |
| 104 | 164 |

***Tablo 3: RandomForest Confusion Matrix***

Yapılan karşılaştırmalardan sonra select attributes bölümünden önerilen özelliklere göre yeniden sınıflandırma yapılmıştır. Önerilen özellikler; plas, mass, pedi ve age özellikleridir. Seçilen özelliklere göre 2 algoritmanın metriklere göre sınıflandırmaları tablo 4’te verilmiştir.

***Tablo 4: Algoritma ve Metrik Karşılaştırmaları( Select Attributes Yapıldıktan Sonra)***

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Metrik  Algoritma | CC | MAE | RMSE | RAE | RRSE | ROC |
| BayesNet | 75.5208 % | 0.3122 | 0.4151 | 68.6807 % | 87.018 % | 0.802 |
| RandomForest | 74.7396 % | 0.3065 | 0.4113 | 67.4438 % | 86.2935 % | 0.813 |

Tablo 4’te görüldüğü gibi select attributes kısmından belirlenen özelliklerden sonra ilk durumun aksine BayesNet doğruluk yüzdesi RandomForest’in doğruluk yüzdesine göre daha iyi olmuştur. Son olarak yeni durumun confusion matrixleri tablo 5 ve 6’da verilmiştir.

|  |  |
| --- | --- |
| a | b |
| 426 | 74 |
| 114 | 154 |

***Tablo 5: BayesNet Confusion Matrix***

|  |  |
| --- | --- |
| a | b |
| 415 | 85 |
| 109 | 159 |

***Tablo 6: RandomForest Confusion Matrix***

4. Değerlendirme

Makine öğrenmesi algoritmalarının aynı veri seti üzerinde karşılaştırılmasıyla istenilen metriklerin değerleri elde edilmiş ve tablolarda her iki durum için karşılaştırılmıştır. Bu durum select attributes yöntemidir. Bu çalışma için kullanılan algoritmalar; BayesNet ve RandomForest algoritmalarıdır. Kullanılan veri seti ise Diyabet veri setidir.

5. Referanslar

[1] <http://www.endustri40.com/makine-ogrenimi-nedir/>

[2] <http://bilgisayarkavramlari.sadievrenseker.com/2009/06/01/weka/>

[3] Kharche, Deepali, K. Rajeswari, and Deepa Abin. "Comparison of different datasets using various classification techniques with weka." *International Journal of Computer Science and Mobile Computing* 3.4 (2014): 389-393.

[4] Aygün, Sercan, and Ece Olcay Güneş. "A benchmarking: Feature extraction and classification of agricultural textures using LBP, GLCM, RBO, Neural Networks, k-NN, and random forest." *Agro-Geoinformatics, 2017 6th International Conference on*. IEEE, 2017.