**“PİMA KIZILDERİLİLERİ DİYABET VERİTABANI”’NIN J48, KNN VE ONER ALGORİTMASI İLE DEĞERLENDİRİLMESİ**

Kenan ÜLKER

1711404043

**ÖZET**

Tıbbi veriler ile ilgili tahmin yapmak amacıyla veri madenciliği analitik araçlarından biri olan sinir ağları kullanılmaktadır. Bir sinir ağı için model seçimi, optimal sayıda gizli düğüm seçimi, ilgili girdi değişkenlerinin seçimi ve optimal bağlantı ağırlıklarının seçimi gibi çeşitli faktörleri gerektirir.

Bu makalemizde Pima Kızılderilileri Diyabet Veritabanı “**“**Pima Indians Diabetes Database**”** örneği kullanılarak Kızılderililerin gebelik sayısı, glikoz, Tansiyon, Deri Kalınlığı, insulün, BMI, soy kütüğü diyabet varlığı, yaş değişkenleri göz önüne alınarak makine öğrenmesi ile kişilerin diyabet hastası olup olmadığını tahmin edeceğiz.

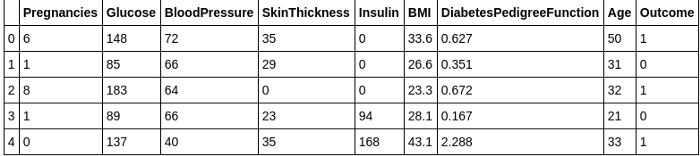
Bu çalışmamızda Veritabanı örneğinde yer alan kişilerin diyabet hastası olup olmadığı konusunda tahminler yapılması amacıyla Weka programı ile analizler yapılmıştır.

**GİRİŞ**

Günümüzde hastanelerde bilgisayarların yağın olarak kullanılması nedeniyle kişiler hakkında çok fazla sayıda veri toplanmaktadır. İnsanların karar vermesi çoğu zaman optimal bir seçim olarak karşımıza çıksa da sınıflandırılacak çok büyük miktarda veri olduğunda yetersiz kalabilmektedir. Tıbbi veri madenciliği, tıp alanının veri kümelerindeki gizli kalıpları keşfetmek için büyük bir potansiyele sahiptir. Bu veri tabanındaki değerlerin kontrolü sağlanarak klinik tanı aşamasında yardımcı olarak kullanılabilir. Sinir Ağları, tıbbi veriler için tahminler yapmak için kullanılabilecek birçok veri madenciliği analitik araçlarından biridir.

**Veri Seti Hakkında**

Bu veri seti; en az 21 yaşında, kadın ve Pima Indian ırkından insanlardan oluşur. Bu insanların verilen belirli özellikleri ile bu kişilerin diyabet hastası olup olmadığını tahminleyeceğiz.



**Tablo 1. Veri Seti İçerik Tablosu**

**Veri Seti İçeriği**

Veri setimizde yer alan değişkenler ve anlamları aşağıda ifade edilmiştir.

Pregnancies = Hamile kalma sayısı

Glucose = Glikoz

Blood Pressure = Kan Basıncı

Skin Thickness = Deri Kalınlığı

Insulin = İnsülin

BMI (Body Mass Index) = Beden Kitle Endeksi

Diabetes Pedigree Function = Soyumuzdaki kişilere göre diyabet olma ihtimalimizi hesaplayan bir fonksiyon

Age = Yaş

Outcome = Diyabet olup olmadığı bilgisi

Bu veri setinde yer alan değişkenlere göre hedefimiz 0 ve 1 değeri olarak kişinin diyabet olup olmadığını tahmin etmeye çalışacağız. Burada 0 diyabet olmaması durumu, 1 ise kişinin diyabet olması durumu olarak değerlendirilmektedir. Veri setinde böylece 2 adet sınıfın yer aldığı da görülmektedir.

Veri setinin incelemesine geçemden önce veri setinin içerdiği bilgilerin ilk bakışta hatalı olup olmadığı konusunda bir inceleme yaptığımızda önemli bir eksikliğin farkına varılmıştır. Üzerinde çalıştığımız “Pima Indians Diabetes Database” veri setindeki 0 değerlerinin tamamının aslında 0 değerini ifade etmediğidir. Burada veri setini girilirken belirlenemeyen değerlere 0 değerinin verildiği düşünülmektedir.

Şöyle ki; Hamile kalma sayısı ve soyumuzdaki kişilere göre diyabet olma ihtimalimizi hesaplayan bir fonksiyon değeri 0 olabilir ancak diğer değişkenlerin glikoz, kan basıncı, deri kalınlığı, beden kitle indeksi vb. 0 olması normal bir durum değildir.

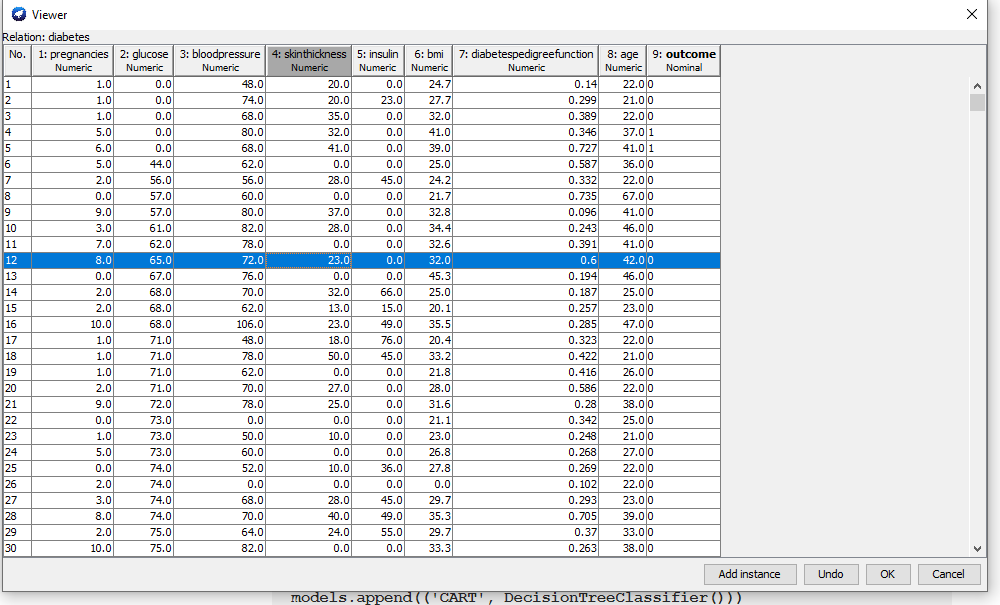
Bu çalışmada hazır veri setinden yararlanılmıştır. “Pima Indians Diabetes Database” üzerinde bilgileri verilen kişilerin diyabet olup olmadığını tahminlemek amacıyla oluşturulan veri seti; aslen Ulusal Diyabet ve Sindirim ve Böbrek Hastalıkları Enstitüsü'nden alınmıştır. Veri setinin amacı, veri setinde yer alan belirli tanısal ölçümlere dayalı olarak bir hastanın diyabetli olup olmadığını tanısal olarak tahmin etmektir.

Veriler ile ilgili kategorik bilgiler Tablo 1’de verilmiştir.

Tablo 1: Veri Setine Ait Bilgiler

|  |  |
| --- | --- |
| Kişi Sayısı | 768 |
| Değişken sayısı | 9 |
| Veri Özelliği | Kategorik Veri |
| Veri işlem | Sınıflandırma |

**Tablo 2. Veri Bilgileri**



**Veri Seti WEKA Görüntüsü**

Tablo 2 incelendiğinde, veri setinin 9 değişken ve 768 kişiden alınan bilgilerden oluştuğu görülmektedir. Veriler kategorik özelliktedir ve verilere sınıflandırma işlemleri yapılmıştır. Veri setinde yukarıda belirtildiği gibi kayıp verilerin olduğu görülmektedir.

**Veri Setinin J48 (Diğer Adıyla C4.5) Karar Ağacının Weka'da Analizi**

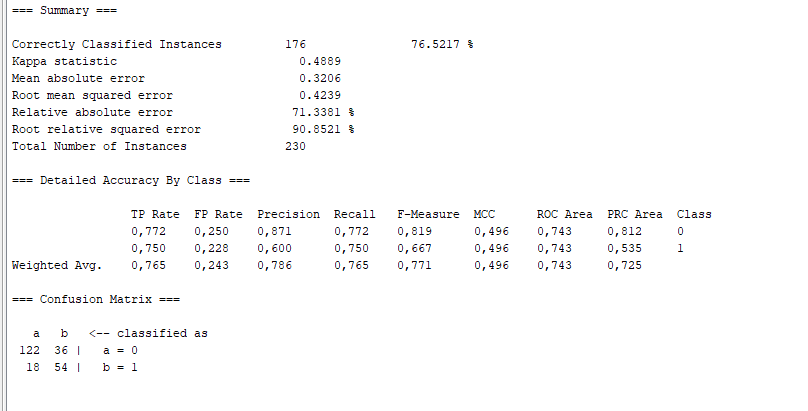
Veri madenciliği uygulamalarında tahmin etme ve tanımlama modelleri kullanılmaktadır. Tahmin edici modeller, sonuçları belli olan verilerden bir model geliştirilip daha sonra sonucu belirli olmayan verilerin sonuçlarının öngörülmesi hedeflenmektedir (Emekci ve ark. , 2007).

|  |  |
| --- | --- |
| Toplam örnek sayısı | 230 |
| Çapraz doğrulama kıvrımları | 9 |
| Yaprak sayısı | 20 |
| İlk dallanma | Glucose |
| Ağaç genişliği | 39 |
| Analiz süresi | 0.04 |
| Doğru sınıflandırılmış örnek sayısı | 176 |
| Doğru sınıflandırma oranı | %76.5217 |
| Kappa istatistiği | 0.4889 |
| Ortalama mutlak hata | 0.3206 |
| Karekök ortalama hata | 0.4239 |

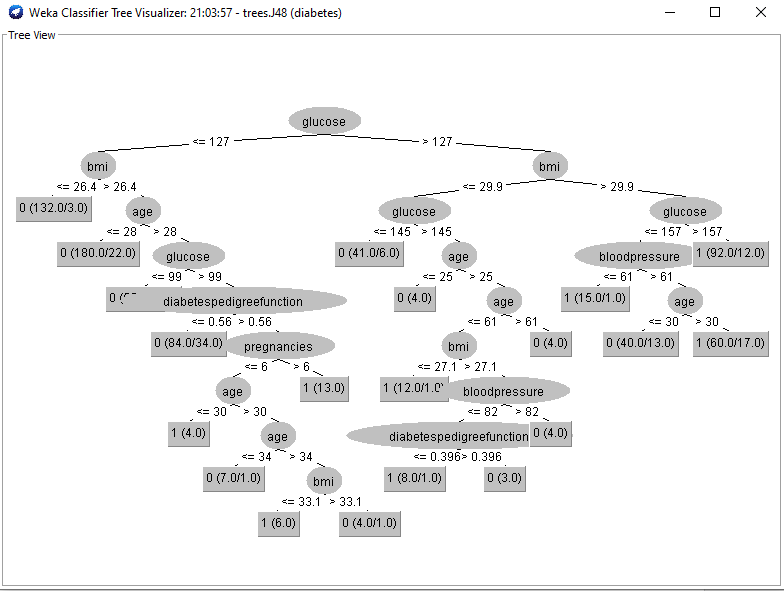
**Tablo 3. Analiz Sonuçları**

WEKA’DA analiz edilen J48 algoritmasına göre ilk dallanmanın glikoz değerinde olduğu belirlenmiştir. Toplam 230 adet örnek üzerinde analiz gerçekleştirilmiştir. % 70 oranında makinenin öğrenmesi sağlanarak kalan kısmının tahmin edilmesi incelenmiştir. Bu veriler doğrultusunda doğruluk oranı % 76,52 olarak gerçekleşmiştir.

Analiz sonuçlarının özeti aşağıda verilmektedir.



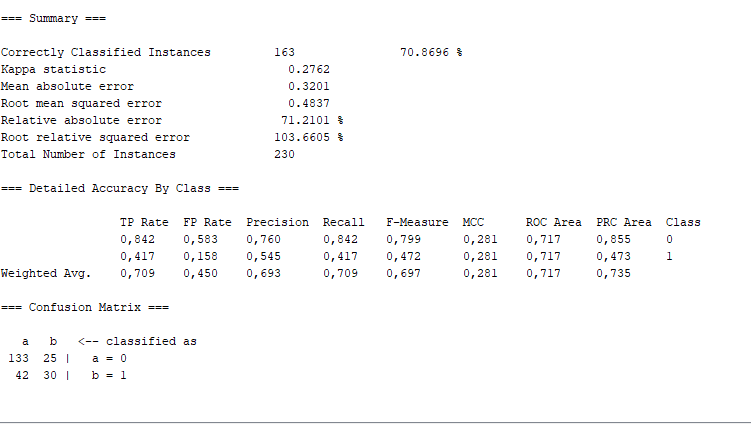
Ağaç yapısı aşağıda verilmiştir.



Karışıklık Matrisi incelendiğinde 122 veri diyabet olması doğru olarak tahmin edilmiş 36 veri ise yanlış tahmin ile diyabetli olarak tahmin edilmiştir. Diğer taraftan 54 kişinin diyabet olması doğru olarak tahmin edilmişken toplam 18 kişinin diyabet olmaması durumu yanlış tahmin edilmiştir. Bu durumda algoritma diyabet konusundaki belirlemede zayıf kaldığı değerlendirilmiştir. Bu değerler ile öğrenme yüzdesi % 80 ve % 90 olarak denenmiş yine doğru tahmin oranı % 75 civarında kalmıştır.

**KNN , K En Yakın Komşu (K Nearest Neighborhood) Algoritması WEKA’da İncelenmesi**

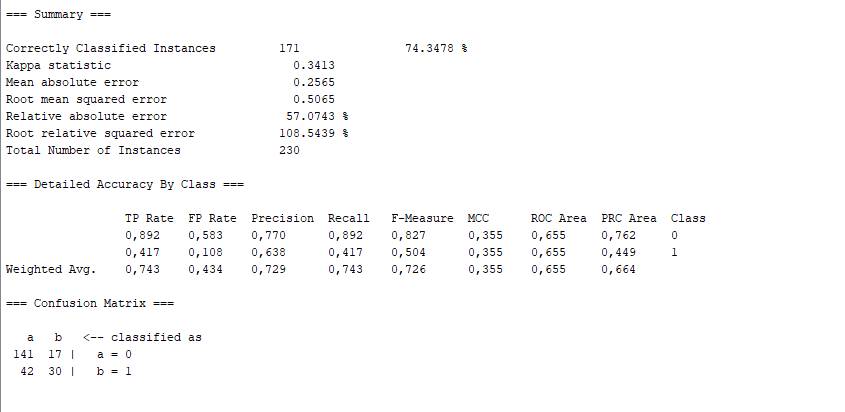
Bu algoritma ile yapılan veri analizlerinde de %70 oranı dikkate alınmıştır. Burada yine 230 adet veri analiz edilmiştir. Analiz sonuçları aşağıda verilmektedir.



Yukarıda verilen görselde görüleceği üzere 133 veri doğru tahmin edilirken 25 veri hatalı olarak tahmin edilmiştir. Makine öğrenme oranı %80’ çıkarıldığında % 72 düzeyinde, %90’ çıkarıldığında ise % 76 gibi bir doğruluk oranına ulaşmaktadır.

**OneR (Tek Sınıf Sınıflandırma ) Yöntemi**

Bu algoritma bir sınıfa göre sonucun belirlenmesi işlemidir.



OneR yönetimi ile veri seti analiz edildiğinde %74 gibi bir oranda doğru tahmin yaptığı görülmüştür. Bu algoritmada 141 kişinin diyabet olma olasılığı doğru 17 kişinin ise yanlış tahmin edilmiştir.

Verilerin makine öğrenme oranı %80 ve %90 seviyelerinde de test edildiğinde bu algoritmada doğruluk oranı %1 dolaylarında arttığı görülmüştür.

**SONUÇ**

Genel olarak bakıldığında makalemizde Pima Kızılderilileri Diyabet Veritabanı “**“**Pima Indians Diabetes Database**”** örneği kullanılarak Kızılderililerin gebelik sayısı, glikoz, Tansiyon, Deri Kalınlığı, insülin, BMI, soy kütüğü diyabet varlığı, yaş değişkenleri göz önüne alınarak makine öğrenmesi ile kişilerin diyabet hastası olup olmadığını tahmin edilmiştir. J48, KNN ve OneR algoritmaları kullanılarak tahminler yapılmıştır. Yapılan tahminlerde %70 oranında makineye öğretilen verilerin %30’luk kısmının tahmin edilmesi amaçlanmıştır. J48’de %76 seviyelerinde olan doğruluk oranı KNN algoritmasında %71 oranında, OneR yönteminde ise %74 oranında doğru tahmin ettiği görülmüştür. Bu incelemeler ile en doğru tahmin işleminin %76 oranı ile J48 algoritması ile yapıldığı değerlendirilmektedir. Ayrıca makine öğrenme oranı %80 ve %90 seviyelerine çıkarılarak analizler yapıldığında doğru tahmin oranında %1-2 dolaylarında düşük bir artışın olduğu gözlemlenmiştir. 4 ayrı algoritmanın doğru olarak yapması gereken tahmin oranının %80’lerin altında kalmasının nedeninin veri kümesinde boş gelen değerlerin 0 olarak dikkate alınmış olması olabileceği değerlendirilmektedir.

**KAYNAKLAR**

Applıcatıon Of Genetıc Algorıthm Optımızed Neural Network Connectıon Weıghts For Medıcal Dıagnosıs Of Pıma Indıans Dıabetes, <https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.206.1621&rep=rep1&type=pdf>

Emekci, F., Sahin, O. D., Agrawal, ve D., El Abbadi, (2007). Privacy preserving decision tree learning over multiple parties, Data & Knowledge Engineering, 63, 348-361.

Makine Öğrenmesi Uygulaması Geliştirme — 1 ( Arff ve Weka), <https://medium.com/@EbubekirBbr/makine-öğrenmesi-uygulaması-geliştirme-1-arff-ve-weka-dd94d274e644>

WEKA, <https://sourceforge.net/projects/weka/>

Pima Indians Diabetes Database, <https://www.kaggle.com/uciml/pima-indians-diabetes-database>