**Makine Öğrenmesi Yöntemleri İle Obezite Seviyelerinin Sınıflandırılması**

**1-) GİRİŞ**

Günümüzde obezite insan sağlığı için önemli bir sorundur.Obezite için önemli noktalar beslenme durumu ve fiziksel durumdur. Bu çalışmada eldeki veriler ve onlara ait öznitelikler ile insanların obezite durumlarına göre sınıflandırma(classfication) işlemi yapılmıştır. Obezite durumları şunlardır ; InsufficientWeight , NormalWeight , ObesityType\_1 ,ObesityType\_2 ,ObesityType\_3 , OverwightLevel\_1 ,OverwightLevel\_2 . Sınıflandırma işleminde kullanılan modeller ise şunlardır ; Naive Bayes , Destek Vektör Makineleri, Rastgele Orman (Random Forest) , k-NN , Karar Ağaçları (Decision Tree) ve Lojistic Regresyon kullanılmıştır.

**2-) Veri Seti**

**Bu çalışmada kullanılan veri seti UCI ‘ den alınmıştır.Veri setinden 17 öznitelik ve 2111 tane veri vardır.Veri setine ait öznitelikler aşağıdaki tabloda gözükmekterdir.**

|  |  |
| --- | --- |
| **Öznitekik** | **Açıklama** |
| **Gender** | **Cinsiyet (Female / Male)** |
| **Age** | **Yaş** |
| **Height** | **Boyu** |
| **Weight** | **Ağırlığı** |
| **family\_history\_with\_overweight** | **Aile üyelerinde aşırı biri varmı (Yes / No )** |
| **FAVC** | **Sık sık yüksek kalorili yiyeceklerinin yenilip yenilmemesi (Yes /No)** |
| **FCVC** | **Öğünlerde sebzeli yemek yenilip yenilmediği (Never ,Sometimes,Always )** |
| **NCP** | **Günde kaç ana öğün yediği (Between 1-2 (0), Three (1), More than Three (2) )** |
| **CAEC** | **Öğün aralarında yemek yenip yenmediği(Always, Frequently, Sometimes, No )** |
| **SMOKE** | **Sigara içip içmediği (Yes / NO )** |
| **CH2O** | **Günden ne kadar su içtiği (Less than a litter (0) , Between 1-2 (1),More than two litter (2) )** |
| **SCC** | **Günlük yenilen kaloriler takip edilip edilmediği (Yes / No )** |
| **FAF** | **Ne sıklıkla fiziksel aktive yapıldığı (No, 1 or 2 days, 2 or 4 days,4 or 5 days)** |
| **TUE** | **Teknolojik aletlerin ne kadar kullanıldığı** |
| **CALC** | **Ne sıklıkla alkol alındığı (No,Sometimes,Frequently,Always )** |
| **MTRANS** | **Genellikle kullanılan ulaşım aracı (Automobile,Bike,Motorbike,PublicTransportain,Walking )** |
| **NObeyesdad** | **Obezite durumu (** InsufficientWeight , NormalWeight , ObesityType\_1 ,ObesityType\_2 ,ObesityType\_3 , OverwightLevel\_1 ,OverwightLevel\_2 **)** |

**3-) Kullanılan Sınıflandırma Algoritmalarının Sonuçları Ve Sonuçların Tartışılması**

**KNN (K-Nearest Neighbors) Algoritması : Tembel bir öğrenme türüdür.KNN temelde.KNN ile temelde bulunan yeni noktaya en yakın noktalar aranır.K , bilinmeyen noktanın yakın komşularının miktarını temsil eder.Yani bir nesne komşuları arasında en yaygın sınıfa verilir.KNN için uzaklık (distance ) hesaplanırken Minkowski uzaklık hesaplama fonksiyonu kullanılır.**

**Rastgele Orman (Random Forest) : Rastgele Orman aslında overfittingin (aşırı öğrenme) önüne geçmek için kullanılır.Rastgele Orman modeli problemi çözmek için hem veri setinde hemde öznitelikleri rastgele olarak 10 larca veya 100 lerce alt set seçip bunları eğitiyior.Bu yöntemle 100 lerce karar ağacı oluşuyuor ve her karar ağacı biriysel olarak tahminde bulunuyuor.En sonunda bu ağaçların ortalması alınıp ona göre sınıflandırma işlemi yapılıyıor.**

**Lojistic Regresyon : Bir sonucu belirleyen bir veya daha fazla bağımsız değişken bulunan bir veri kümesini analiz etmek için kullanılan bir yöntemdir.**

**Naive Bayes :** **Naive Bayes sınıflandırıcısının temeli Bayes teoremine dayanır. Tembel bir öğrenme algoritmasıdır aynı zamanda dengesiz veri kümelerinde de çalışabilir. Algoritmanın çalışma şekli bir eleman için her durumun olasılığını hesaplar ve olasılık değeri en yüksek olana göre sınıflandırır.**

**Karar Ağaçları (Decision Tree) :** **En basit tanım ile, olası sonuç için oluşturulabilecek karar yollarının oluşturulmasıdır. Elimizdeki veri setindeki bir öznitelik seçip onun diğer alanlar üzerindeki dallanmalarını ilişkilendirip bir karar ağacı çıkartırız.Entropi paremetresi, rastgele bir değişkenin belirsizliğinin ölçüsüdür.**

**Desktek Vektör Makineleri :** **Destek vektör makineleriyle sınıflandırmada, iki sınıfa ait örneklerin doğrusal olarak dağıldığını varsayalım. Bu durumda bu iki sınıfın, eğitim verisi kullanılarak elde edilen bir karar fonksiyonu yardımıyla birbirinden ayrılması amaçlanır.**

**Kullanılan sınıflandırma algoritmalarında sonuçların daha iyi incelenebilmesi için confusion matris üzerinden accuracy , f1-score, recall, precision, roc eğrisin altında kalan alan (auc) değerleri bulunmuştur.Bu değerler 10 kat çapraz doğrulama test tekniğine göre bulunmuştur.**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **Accuracy** | **F1-score** | **Recall** | **Precision** | **Auc** |
| **k-NN** | **0,69314** | **0,67201** | **0,69314** | **0,68469** | **0,90044** |
| **Random Forest** | **0,81825** | **0,83259** | **0,83904** | **0,86305** | **0.97280** |
| **Lojistic Regresyon** | **0,78988** | **0,78239** | **0,78988** | **0,79681** | **0,95636** |
| **Naive Bayes** | **0,51506** | **0,46600** | **0,51506** | **0,48960** | **0,81634** |
| **Karar Ağaçları** | **0,87710** | **0,88268** | **0,88088** | **0,89780** | **0,93345** |
| **Destek Vektör Makineleri** | **0,89583** | **0,89368** | **0,89583** | **0,90621** | **0,98679** |

**Accuracy : Modelde doğru tahmin edilen verilerin tüm verilere oranıdır.**

**F1-score : Precision ve Recall değerlerinin harmonik ortalmasıdır.**

**Precision : Doğru olarak tahmin ettimiz verilerin yüzde kaçını gerçekten doğru olduğudur.**

**Recall : Doğru olarak tahmin edilen verilerin gerçekte yüz kaçının doğru olduğudur.Algoritmanın doğru tahmin etmesinin bağımsız gerçekteki doğrular (TP+FN).**

**Auc: Roc eğrisinin altında kalan alandır.**

**Tablodaki verilen değerlere göre veri setimiz için en iyi sonucu veren algoritmalar sırasıyla şöyledir (AUC Değerine Göre Sıralanmıştır ):**

**1-) Destek Vektöre Makineleri**

**2-) Random Forest**

**3-) Lojistic Regresyon**

**4-) Karar Ağaçları**

**5-) k-NN**

**6-) Naive Bayes**

**Sonuç olarak bu veri seti için kullanılan modellerle bazıları yüksek bir yüzde verip doğru sınıflandırma yapmıştır bazıları ise yetersiz kalmıştır.**

**KAYNAK DOSYALAR**

**Bu çalışmada kullanılan veri setine https://archive.ics.uci.edu/ml/datasets/Estimation+of+obesity+levels+based+on+eating+habits+and+physical+condition+ adresinden ulaşılabilir.**