**1.Dünya Ülkelerinin Sınıflandırma Algoritmaları ile Tahmin Edilmesi**

Elif Karagöz , Halit Kaan Kaçar

Yeditepe Üniversitesi, Uygulamalı Bilimler Yüksekokulu, Yazılım Geliştime Bölümü, İstanbul, Türkiye, elif.karagoz@yeditepe.edu.tr

Yeditepe Üniversitesi, Uygulamalı Bilimler Yüksekokulu, Yazılım Geliştime Bölümü, İstanbul, Türkiye, halit.kacar@yeditepe.edu.tr

**Öz**

Verilen veri seti üzerinden 1. Dünya Ülkeleri tahmini yaptırılması amaçlanmıştır. Bu çalışmada öncelikli olarak “None” değerler o grubun ortalama değeri ile doldurulmuş, sonrasında veri setindeki öznitelik grupları kendi aralarında karşılaştırılıp grafiklendirilme yapılmıştır. Doğrusal Regresyon kullanılarak medyan yaşa göre doğurganlık oranı çıkarımı yaptırılmıştır. Hedef 1. Dünya Ülkeleri tahmini için veri setinin yüzde otuzu test için ayrılmış, random forest skoru hesaplatılmış ve sonrasında Karışıklık Matrisi kullanılarak verinin doğruluğu hesaplanmıştır.

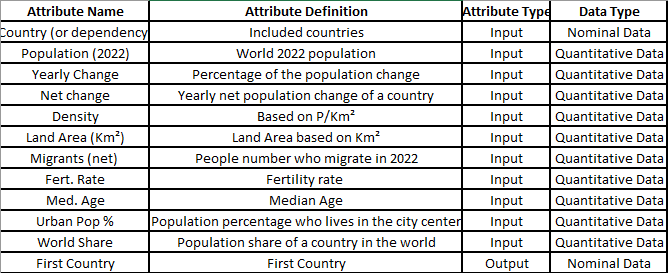
**1. Giriş**

Dünya nüfusu her geçen gün katlanarak artmaktadır. En büyük nüfusa sahip ülke Çin olmakla beraber en az nüfusa sahip ülke Holy See’dir. Bu çalışmada tek tek ülkerin 2021-2022 yılları arasındandaki nüfus artışı ve dağılımı incelenmiştir. İnceleme yapılırken, seçilen ülkelerin doğurganlık oranına kıyasla medyan yaşı ve o ülkenin merkezi nüfusu gibi verilerle kıyaslama yapılmıştır. Bu bulgular doğrultusunda ülkelerin gelişmiş ülke olup olmadığının tahmin edlmesi hedeflenmiştir.

**2. Materyal ve Metot**

Bu çalışmada Kaggle web sitesinden alınan 2022 dünya popülasyon artış veri seti kullanılmıştır. Veri setinde toplam 12 öznitelik ve 243 ülkeye ait popülasyon veri kayıtlarını içermektedir. Öznitelik değerleri ve açıklamaları Tablo 1’de gösterilmiştir.

*Tablo 1. Öznitelik Tablosu*

****

**2.1. Öznitelik Seçimi**

Çalışmada öznitelik değerleri seçilirken öznitelikler arası bağlantı incelenmiş ve buna göre mantıksal bağlantısı en fazla olan alanlar kendi aralarında ikişerli olmak üzere karşılaştırılmış ve tahmin fonksiyonuna tabi tutulmuştur.

Linear regresyon algoritmasından yararlanılarak iki alan arası ilişki incelenip tahmin doğrusu oluşturulmuştur. Lineer (Doğrusal) regresyon istatistiksel veri analizinde doğrusal ve sürekli değişkenler için sıkça kullanılan, bir yöntemdir. Tek bir bağımsız değişken olduğunda regresyon modeline Basit Doğrusal Regresyon, birden fazla bağımsız değişken olduğunda ise Çoklu Doğrusal Regresyon Modeli denir. Çalışmamızda medyan yaş ile doğurganlık oranını modellemek için Basit Doğrusal Regresyon modeli kullanılmıştır.

**2.2. Sınıflandırma Algoritmaları**

1.Dünya Ülkelerinin tahmini için gerçekleştirilen bu çalışmada NaN değerlerinin doldurulması aşamasında K-En Yakın Komşu (KNN, K-nearest neighborhood), medyan yaş ve doğurganlık oranı arasındaki bağıntıyı gözlemleyip tahmin modeli oluşturmak adına Linear Regresyon sınıflandırma algoritmaları kullanılmıştır.

***2.2.1 K-En Yakın Komşu (KNN)***

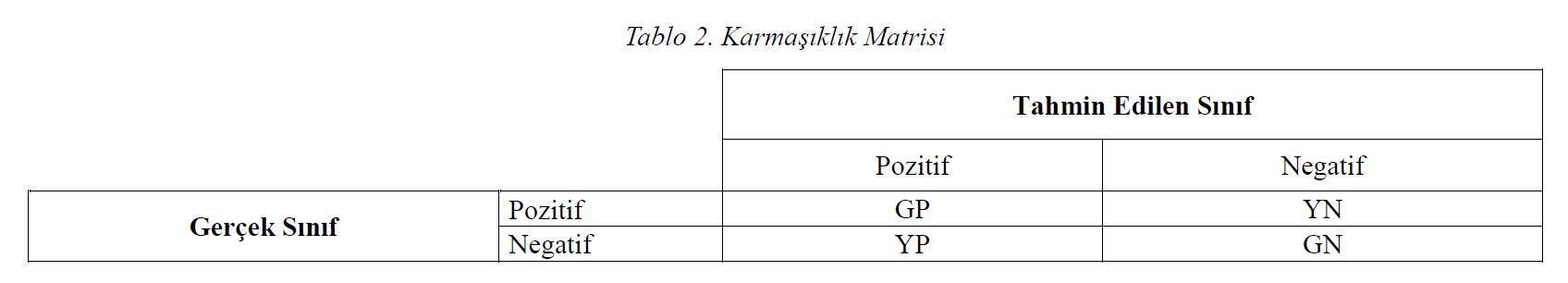
En Yakın Komşu Algoritması (K-Nearest Neighbours – KNN); sınıflandırma işleminde bulunulacak örnek veri noktasının bulunduğu sınıfın (öğrenim kümesi) ve en yakın komşunun (elemanın), k değerine (benzerliğe) göre belirlendiği bir denetimli/gözetimli makine öğrenme yöntemi olarak ifade edilmektedir. Algoritmaya yönelik uzaklık hesaplaması tipleri Euclidean, Manhattan, Chebyshev, Hamming, Minkowski ve dahası olmak üzere çok çeşitlidir.

***2.2.2 Linear Regresyon***

Doğrusal regresyon analizi, bir değişkenin değerini başka bir değişkenin değerine göre tahmin etmek için kullanılır. Tahmin etmek istediğiniz değişken, bağımlı değişken olarak adlandırılır. Diğer değişkenin değerini tahmin etmek için kullandığınız değişken ise bağımsız değişken olarak adlandırılır.

Bu analiz biçimi, bağımlı değişkenin değerini en iyi öngören bir ya da daha fazla bağımsız değişkeni kullanarak doğrusal denklemin katsayılarını tahmin eder. Doğrusal regresyon, öngörülen ve gerçek çıkış değerleri arasındaki uyumsuzlukları en aza indiren düz bir çizgi ya da yüzeye yerleşir. Bir çift eşleştirilmiş veri kümesi için en uygun satırı keşfetmek üzere "en küçük kareler" yöntemini kullanan basit doğrusal regresyon hesaplayıcılar vardır. Daha sonra, Y'den (bağımsız değişken) X'in (bağımlı değişken) değeri tahmin edilir.

**2.3. Değerlendirme Kriterleri**

Eğitim veri seti ile sınıflandırma modelleri oluşturulduktan sonra test verisi kullanılarak sınıflandırıcıların performansları test edilmektedir. Sınıflandırıcıların performanslarının test edilebilmesi için çeşitli değerlendirme kriterleri hesaplanmaktadır. Bu çalışmada değerlendirme kriterlerinin hesaplanması için karmaşıklık matrisinden yararlanılmıştır. Karmaşıklık Matrisi Tablo 2’de verilmiştir. 

Gerçek Pozitif (GP), gerçek sınıfı pozitif olan bir örneğin doğru sınıflandırıldığı durumları göstermektedir.

Yanlış Negatif (YN), gerçek sınıfı pozitif olan bir örneğin yanlış sınıflandırıldığı durumları göstermektedir.

Yanlış Pozitif (YP), gerçek sınıfı negatif olan bir örneğin yanlış sınıflanırıldığı durumları göstermektedir.

Gerçek negatif (GN), gerçek sınıfı negatif olan bir örneğin doğru sınıflandırıldığı durumları göstermektedir.

Karmaşıklık matrisinin hücrelerindeki değerler kullanılarak doğru sınıflandırma oranı, kesinlik, duyarlılık, F ölçütü ve Kappa istatistiği değerlendirme ölçütleri hesaplanabilmektedir.

Doğru sınıflandırma oranı, doğru sınıflandırılan örneklerin tüm örneklere oranı ile bulunmaktadır. Formülasyonu aşağıdaki eşitlikte verilmiştir.

Doğru Sınıflandırma Oranı=(GP+GN)/(GP+YN+YP+GN)

Kesinlik, pozitif olan ve doğru tahmin edilen örneklerin pozitif olarak tahmin edilen örneklerin toplamına oranıdır. Eşitlik (4)’teki gibi hesaplanmaktadır.

Kesinlik=GP/(GP+YP)

Duyarlılık, pozitif olan ve doğru tahmin edilen örneklerin pozitif örneklerin toplamına oranıdır.

Duyarlılık=GP/GP+YN

F ölçütü keskinlik ve duyarlılık değerlerinin harmonik ortalaması alınarak hesaplanmaktadır. F ölçütünün hespalanmasına ilişkin formül aşağıda verilmiştir.

F ölçütü=(2 x Duyarlılık x Keskinlik)/(1-Beklenen Doğruluk)

**3.Araştırma Sonuçları ve Tartışma**

Bu çalışmada dünya popülasyonu verisi kullanılarak K-En Yakın Komşu, Linear Regresyon sınıflandırıcılarının performansları değerlendirilmiş ve 1.Dünya ülkeleri tahmin ettirilmiştir. Bu sonuca giderken dağılım grafiği üzerinde en uygun veri türleri arasında karşılaştırma yapılırken doğurganlık oranı ve medyan yaş arasındaki bağıntı incelenmiş y = e^x bağıntısı gözlemlenmiş, kentsel popülasyon ve doğurganlık oranı arasındaki ilişki incelenirken, kentselde yaşayan kesimin doğruganlık oranının daha düşük olduğu fakat, kırsal kesimde yaşayan insanların doğurganlık oranının daha yüksek gözlemlenmiş ve bu ilişki eğitim seviyesi farkı ile yorumlanmıştır. Pasta grafiklerinde ise yoğunluğu (popülasyon/ülke yüzey km2), ülke popülasyonu ve dünya üzerindeki payı en çok olan ilk 10 ülke incelenirken ortaya Çin’in en çok populasyona ve yüzey alanına sahip olduğu fakat km2 başına düşen insan sayısının diğer çoğu ülkeye kıyasla düşük olduğunu göstermiştir. Bu da ülkelerin gelişmiş olup olmadığını popülasyondan bulmaya çalışmanın yanlış olduğunu, bunun yerine doğum oranı ve yüzey alanına düşen kişi sayısı gibi verilerden sağlamanın daha sağlıklı olduğunu kanıtlar niteliktedir.

**4.Sonuç**

Çalışmanın sonucunda değerlendirme ölçütlerine göre en başarılı sınıflandırıcının %93 oranına sahip olduğu görülmüştür. 1.Dünya Ülkelerinin %93 oranında doğru tahmin edildiği gözlemlenmiştir. Bu sonuç daha önce belirlemiş olduğumuz ülkelerin; nüfuslarının dağılımı, doğum oranı, göç, nüfus yoğunluğu, ortanca yaş gibi özelliklerle gelişmiş olup olmadığını tahmin edebilme hedefimizin başarılı bir şekilde sonuçlandığını kanıtlar niteliktedir.