**VERİ MADENCİLİĞİ**

Bu proje, COVID-19 verilerini kullanarak ülkelerin ölüm oranlarının yüksek veya düşük olduğunu tahmin etmeye çalışıyor. Bunun için toplam vaka, ölüm, iyileşen, test sayısı ve nüfus gibi özelliklerden faydalandık. Veri temizlendi: sayısal değerlerdeki virgüller kaldırıldı, eksik veriler dolduruldu. **Hedef değişken oluşturuldu: Ölüm sayısı medyan değerinden yüksekse 1, düşükse 0 olarak etiketlendi**. Dört farklı sınıflandırma algoritması kullandık: **Karar** **Ağacı**, **Rastgele Orman**, **K-En Yakın Komşu** ve **Naive Bayes**. Her model eğitim verisiyle öğrenip test verisinde tahmin yaptı.

**Model Doğruluk Karşılaştırması**

| **Model** | **Doğruluk (Accuracy)** | **Notlar** |
| --- | --- | --- |
| **Decision Tree** | **1.0000 (100%)** | Tüm test verisini %100 doğru tahmin etmiş. |
| **Random Forest** | **1.0000 (100%)** | Aynı şekilde %100 doğruluk sağlamış. |
| **KNN** | **0.9574 (95.74%)** | Yüksek doğruluk, ancak 2 örneği yanlış sınıflandırmış. |
| **Naive Bayes** | **0.9362 (93.62%)** | En düşük doğruluk, 3 örnek yanlış tahmin edilmiş. |

**Modellerin Ayrıntılı Yorumu**

**1. Decision Tree**

* **Avantajı**: Basit yapılı ve kolay yorumlanabilir bir modeldir.
* **Doğruluk**: %100 – tüm sınıfları doğru ayırmış.
* **Dikkat**: Bu kadar yüksek doğruluk, aşırı öğrenme (overfitting) ihtimalini gösterir. Eğitim verisini ezberlemiş olabilir.

**2. Random Forest**

* **Avantajı:** Birden fazla karar ağacının ortalaması alınarak daha sağlam tahmin yapılır.
* **Doğruluk:** %100 – oldukça başarılı.
* **Farkı:** Decision Tree’ye göre daha genelgeçer ve overfitting riski daha düşüktür.

**3. K-Nearest Neighbors (KNN)**

* **Doğruluk:** %95,74 – 2 örnek yanlış tahmin edilmiş.
* **Avantajı:** Basit ve güçlüdür, özellikle benzerlik bazlı sınıflandırmalarda iyidir.
* **Dezavantajı:** Veri ölçeklendirmesi şarttır ve büyük veri setlerinde yavaştır.

**4. Naive Bayes**

* **Doğruluk:** %93,62 – 3 yanlış tahmin.
* **Avantajı:** Hızlıdır ve küçük veri setlerinde iyi çalışır.
* **Dezavantajı:** Özelliklerin birbirinden bağımsız olduğunu varsayar, bu her zaman doğru olmayabilir.

**Genel Yorum ve Karşılaştırma**

* **En iyi performansı Decision Tree ve Random Forest gösterdi (%100 doğruluk)**, ancak Decision Tree aşırı öğrenmiş olabilir. Gerçek hayatta bu kadar yüksek doğruluk nadir görülür, genelde uyarıcıdır.
* **KNN ve Naive Bayes de oldukça başarılı**, ama birkaç küçük hata yaptılar.
* **Random Forest**, genellikle en dengeli ve güvenilir yöntemdir hem doğruluk hem de genelleme yeteneği açısından.

**Ayrıntılar**

**1. Veri Seti Hakkında**

* **Veri seti adı:** covid\_worldwide.csv
* **Kaynak:** [**https://www.kaggle.com/datasets/themrityunjaypathak/covid-cases-and-deaths-worldwide/data**](https://www.kaggle.com/datasets/themrityunjaypathak/covid-cases-and-deaths-worldwide/data)
* **İçerik:** Dünya genelindeki ülkelerin pandemi dönemindeki COVID-19 istatistikleri.

**2. Kullanılan Değişkenler**

* **Total Cases:** Toplam COVID-19 vaka sayısı
* **Total Deaths:** Toplam ölüm sayısı
* **Total Recovered:** Toplam iyileşen hasta sayısı
* **Active Cases:** O anki aktif vaka sayısı
* **Total Test:** Toplam yapılan test sayısı
* **Population:** Ülke nüfusu

**Ek olarak:**

* **High\_Death\_Rate:** Modelin tahmin etmeye çalıştığı hedef değişken.
  + **1:** Ülkenin toplam ölüm sayısı medyanın üstünde (yüksek ölüm oranı)
  + **0:** Medyanın altında (düşük ölüm oranı)

**3. Analiz ve Kullanılan Algoritmalar**

**Veri Ön İşleme**

* Sayısal verilerdeki binlik ayırıcı (virgül) ve boşluklar temizlendi.
* Eksik veriler ortalama değerle dolduruldu.
* Gereksiz sütunlar (ülke adı, seri numarası) çıkarıldı.
* Veriler eğitim ve test setlerine ayrıldı (%80 eğitim, %20 test).
* Veriler ölçeklendirildi (StandardScaler ile).

**Kullanılan Algoritmalar**

| **Algoritma** | **Açıklama** |
| --- | --- |
| Decision Tree | Karar ağacı yöntemi, veriyi dallandırarak sınıflandırma yapar. |
| Random Forest | Birçok karar ağacının kombinasyonu ile daha sağlam model oluşturur. |
| K-Nearest Neighbors (KNN) | Komşuluk bazlı sınıflandırma, benzer örneklere göre karar verir. |
| Naive Bayes | Olasılıksal sınıflandırma, özelliklerin bağımsız olduğunu varsayar. |

**Sonuçlar**

| **Model** | **Doğruluk (Accuracy)** | **Yorum** |
| --- | --- | --- |
| **Decision Tree** | **%100** | Mükemmel sonuç, ancak aşırı öğrenme olabilir. |
| **Random Forest** | **%100** | En dengeli ve güvenilir sonuç. |
| **KNN** | **%95,7** | Çok başarılı, birkaç hata yapıyor. |
| **Naive Bayes** | **%93,6** | En düşük doğruluk, ancak hızlı ve pratik. |

**4. Performans Karşılaştırması ve Model Tercihi**

* **Decision Tree** ve **Random Forest** modelleri en yüksek doğruluğu sağladı. Ancak **Decision Tree** tek başına aşırı öğrenme riski taşıdığı için genelleme yeteneği zayıf olabilir.
* **Random Forest** ise birden fazla karar ağacının birleştirilmesiyle daha sağlam ve genellenebilir sonuçlar verdi. Bu nedenle en iyi model olarak tercih edilmiştir.
* **KNN** ve **Naive Bayes** modelleri de başarılı olmakla beraber, doğruluk oranları biraz daha düşük kaldı.
* Sonuç olarak, veri setinin yapısı ve karmaşıklığı göz önüne alındığında **Random Forest** algoritması, yüksek doğruluk ve genellenebilirlik açısından en uygun modeldir.