Yapay Zeka 2022/2

Ödev No:2

Ödev Konusu: Bir veri kümesi oluşturup makine öğrenmesi algoritmalarını çalıştırmak.

Proje Ekibi: Ömer Talha BAYSAN – 18011103

Tolga SAĞLAM – 18011064

Projemizi konusu telefon fiyat tahmin modelidir. Veri seti "ParseHub" uygulaması kullanılarak Epey sitesinden telefon bilgileri çekilerek hazırlanmıştır.

Veri seti özellikleri:

- **1.Phone_name** Cep telefonlarının ve Markaların adını ifade eder.
- **2.Phone_Ratings** Bu Özellik, tüketicilerin her cep telefonu için verdiği puanların sayısını ifade eder.
- **3.Phone_RAM** Telefonun RAM boyutuna sahiptir. Her cep telefonundaki ROM (Dahili Bellek) GB cinsinden miktarını belirtir.
- **4.Phone_ROM** Telefonun ROM (Dahili Bellek) boyutuna sahiptir. Her cep telefonundaki ROM (Dahili Bellek) GB cinsinden miktarını belirtir.
- **5.Phone_Mobile_Size** Belirli bir cep telefonunun kaç inç olduğunu gösterir. Burada tüm değerler inç cinsinden verilmiştir.
- **6.Phone_Primary_Cam** Her cep telefonu için birincil kameranın (Arka Kamera) piksel sayısını belirtir.
- **7.Phone_Selfi_Cam** Her cep telefonu için Selfie kamerasının (Ön Kamera) piksel sayısını belirtir.
- **8.Phone_Battery_Power** Her cep telefonundaki pil gücünün mAh cinsinden miktarını belirtir.
- **9.Phone_Prices** Veri kümesinin Bağımlı Bir Özelliğidir. Sadece her cep telefonunun fiyatlarını gösteriyor.

Model oluşturulmak için Random Forest Regressor, Support Vector Regressor, Gradient Boosting Machines, XGBoost ve Light GBM algoritmaları kullanılmıştır.

Kullanılan algoritmaları karşılaştırmak için "Cross Validation Score" değerleri kullanılmıştır. Bu değerler birbirleri ile karşılaştırılmak için t-testi kullanılmıştır. Alınan skorlar aşağıdaki gibidir.

```
                  Algorithm A                  Algorithm B  T-Statistic
\
0                   Light GBM                      XGBoost    -1.998030
1                   Light GBM  Gradient Boosting Machines    -1.273147
2                   Light GBM    Support Vector Regressor     1.409589
3                   Light GBM     Random Forest Regressor    -1.078186
4                     XGBoost  Gradient Boosting Machines     0.464925
5                     XGBoost    Support Vector Regressor     3.554013
6                     XGBoost     Random Forest Regressor     0.729451
7  Gradient Boosting Machines    Support Vector Regressor     2.625755
8  Gradient Boosting Machines     Random Forest Regressor     0.212077
9    Support Vector Regressor     Random Forest Regressor    -2.456223

    P-Value
0  0.061052
1  0.219163
2  0.175705
3  0.295188
4  0.647560
5  0.002268
6  0.475112
7  0.017144
8  0.834430
9  0.024427
```

'T-Statistic' ve 'P-Value' değerlerini karşılaştırdıktan sonra aşağıdaki adımları izleye biliriz:

1 - T-Statistic değeri: T-istatistiği, iki grup arasındaki farkın istatistiksel olarak ne kadar anla mlı olduğunu ölçer. Pozitif bir T-Statistic değeri, grupların arasında anlamlı bir fark olduğunu gösterirken, negatif bir değer ise bir grup lehine anlamlı bir fark olduğunu gösterebilir. T-Stat istic değeri, büyüklüğüne bağlı olarak değerlendirilebilir. Örneğin, daha yüksek bir T-Statisti c değeri, gruplar arasındaki farkın daha belirgin olduğunu gösterebilir.

2 - P-Value değeri: P-değeri, t-testinin istatistiksel olarak anlamlı olduğunu gösteren bir ölçüd ür. P-değeri, iki grup arasındaki farkın rastgele oluşma olasılığına karşılık gelir. Genellikle, b elirlenen bir anlamlılık düzeyi (örneğin, $p < 0.05$) kullanılarak p-değeri değerlendirilir. Eğer p -değeri, belirlenen anlamlılık düzeyinden küçükse, yani $p < 0.05$ ise, bu durumda iki grup ara sındaki farkın istatistiksel olarak anlamlı olduğu kabul edilir.

Sonuçları değerlendirirken, aşağıdaki genel yaklaşımı kullanabiliriz:

- Eğer T-Statistic değeri pozitif ve p-değeri belirlenen anlamlılık düzeyinden küçükse ($p < 0.05$), o zaman iki grup arasında anlamlı bir fark olduğu ve bir algoritmanın diğeri ne göre daha iyi performans gösterdiği söylenebilir.
- Eğer T-Statistic değeri negatif ve p-değeri belirlenen anlamlılık düzeyinden küçükse ($p < 0.05$), o zaman diğer algoritmanın daha iyi performans gösterdiği ve iki grup arası nda anlamlı bir fark olduğu söylenebilir.

- Eğer p-değeri belirlenen anlamlılık düzeyinden büyükse ($p \geq 0.05$), o zaman istatistiksel olarak anlamlı bir farkın olmadığı ve algoritmalar arasında performans açısından bir fark olmadığı söylenebilir.

Tabii ki, bu sonuçlar yalnızca istatistiksel analizin sonuçlarıdır ve gerçek dünya durumunu tam olarak yansıtmayabilir. Diğer faktörleri ve bağlamı dikkate alarak sonuçları yorumlamanız önemlidir.