BLM4510 Yapay Zeka
Proje Ödevi

Hazırlayanlar:
Onur Demir 18011078
Melih Ocakcı 19011061

Video Adresi:
https://youtu.be/i-Gv4eq5iGA

Konunun Tanıtımı

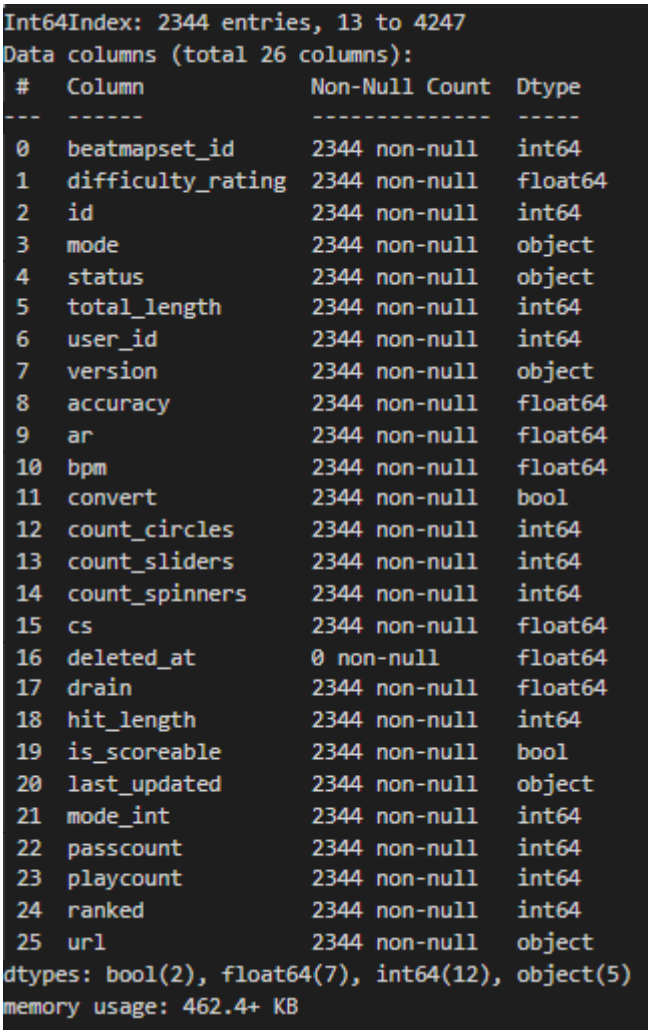
osu! Oyunundaki haritaların verileri internet üzerinden paylaşılmaktadır. Kişiler bu haritaları indirebilmekte ve bu haritaları oyuna yükleyerek haritayı oynamaktadır. Oyunun haritasını indirdikleri sayfa içerisinde aynı zamanda oyuna ait haritanın bilgisi bulunmaktadır. Bu bilgileri kullanarak o haritaya ait zorluğun tahminini yapabilecek model geliştirilmek istenmektedir.

Geliştirme Süreci

Veriyi çekmek için osu! nun internet sitesindeki haritaları inceledik. Sayfadan aşağı indiğimizde her şeferinde gelecek sefer verilecek olan veriyi daha hızlı yüklemesi için json verisi olarak browserda tutulmaktaydı bizde bu veriyi ajax çağrısı kullanarak aldık ve içerisindeki gereksiz veriyi('url','artist adları'…) temizledik.

Veri Seti Tanıtımı ve Haritalar

Aşağıdaki şekilde veri setine ait attribute lar verilmiştir. Biz bu attribute lardan 6 tanesini kullanmayı seçtik ve mode olarak sadece osu! nun yuvarlak bazlı haritalarını alacak şekilde filtereledik.



```
Int64Index: 2344 entries, 13 to 4247
Data columns (total 26 columns):
 #   Column             Non-Null Count  Dtype
---  ------             --------------  -----
 0   beatmapset_id      2344 non-null   int64
 1   difficulty_rating  2344 non-null   float64
 2   id                 2344 non-null   int64
 3   mode               2344 non-null   object
 4   status             2344 non-null   object
 5   total_length       2344 non-null   int64
 6   user_id            2344 non-null   int64
 7   version            2344 non-null   object
 8   accuracy           2344 non-null   float64
 9   ar                 2344 non-null   float64
 10  bpm                2344 non-null   float64
 11  convert            2344 non-null   bool
 12  count_circles      2344 non-null   int64
 13  count_sliders      2344 non-null   int64
 14  count_spinners     2344 non-null   int64
 15  cs                 2344 non-null   float64
 16  deleted_at         0 non-null      float64
 17  drain              2344 non-null   float64
 18  hit_length         2344 non-null   int64
 19  is_scoreable       2344 non-null   bool
 20  last_updated       2344 non-null   object
 21  mode_int           2344 non-null   int64
 22  passcount          2344 non-null   int64
 23  playcount          2344 non-null   int64
 24  ranked             2344 non-null   int64
 25  url                2344 non-null   object
dtypes: bool(2), float64(7), int64(12), object(5)
memory usage: 462.4+ KB
```

Şekil 1: Ajax çağrısı ile çekilen 25 sütunluk veri

Kullanılacak Sütunların Seçimi ve Ön İşleme

Eğitim aşamasında kullanmak için 6 adet özellik seçtik. Bunlar Şekil 2'de görülebilir. Şekil 3'de verilerin normalizasyondan sonraki halleri görülmektedir.

	hit_length	cs	drain	ar	accuracy	bpm
count	2344.000000	2344.000000	2344.000000	2344.000000	2344.000000	2344.000000
mean	151.779010	3.828925	4.683319	7.798422	6.862799	165.457149
std	75.316987	0.634053	1.086388	1.742350	1.919398	37.368717
min	30.000000	0.000000	0.000000	0.000000	0.000000	57.333300
25%	87.000000	3.500000	4.000000	7.000000	6.000000	140.000000
50%	134.500000	4.000000	5.000000	8.500000	7.500000	172.000000
75%	207.000000	4.000000	5.400000	9.000000	8.400000	187.000000
max	625.000000	10.000000	9.000000	10.000000	10.000000	300.999000

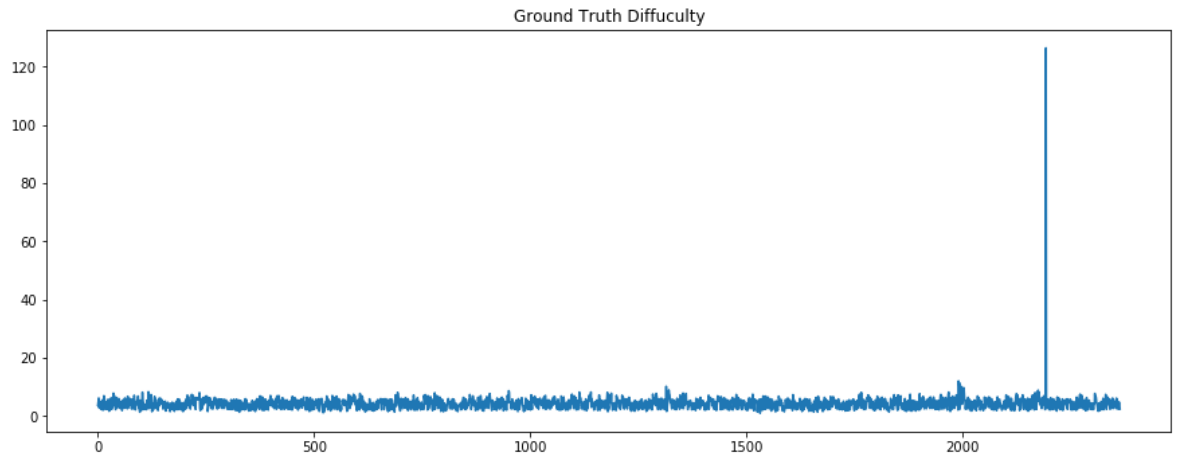
Şekil 2: Normalizasyon işleminden önce

	0	1	2	3	4	5
count	2344.000000	2344.000000	2344.000000	2344.000000	2344.00000	2344.000000
mean	0.204671	0.382892	0.520369	0.779842	0.68628	0.443738
std	0.126583	0.063405	0.120710	0.174235	0.19194	0.153361
min	0.000000	0.000000	0.000000	0.000000	0.00000	0.000000
25%	0.095798	0.350000	0.444444	0.700000	0.60000	0.339263
50%	0.175630	0.400000	0.555556	0.850000	0.75000	0.470590
75%	0.297479	0.400000	0.600000	0.900000	0.84000	0.532150
max	1.000000	1.000000	1.000000	1.000000	1.00000	1.000000

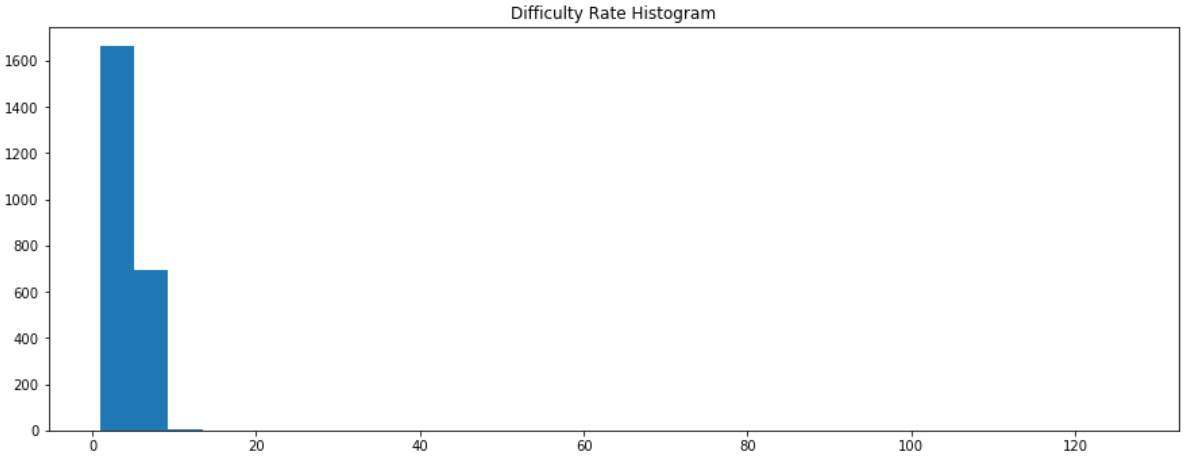
Şekil 3: Normalizasyon işleminden sonra

Tahmin Edilecek Verideki Outlier Tespiti

Verimizi ilk olarak eğitimimizde değişen kesinlik bilgileri ile karşılaştık bunun için verinin label(difficulty) kısmını inceledik. Bu kısımda aşağıdaki değerleri elde ettik 125 zorluklu bir harita bulunmaktadır bu da bizim rank olarak dereceli oyunları değil normal oyunları da aldığımızı göstermektedir. Bu yüzden bu veriyi sadece dereceli oyunları içerecek şekilde kırptık.

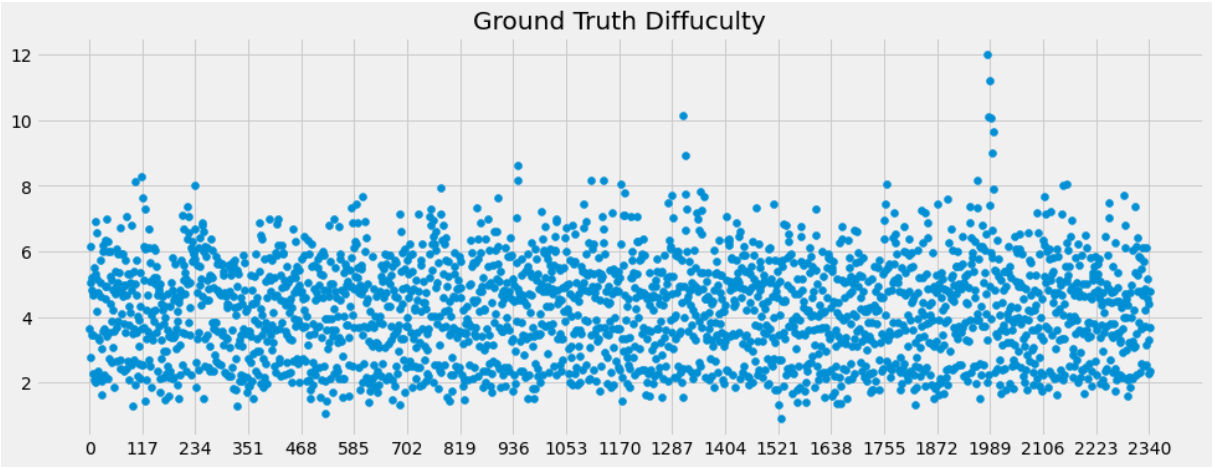


Şekil 4: 125 zorluklu harita model yüzdesini etkilemektedir.

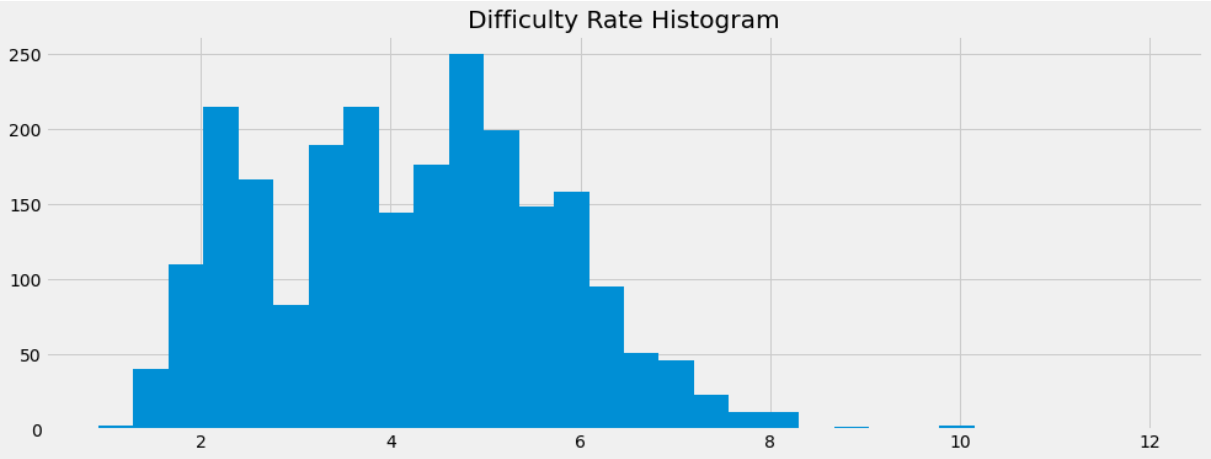


Şekil 5: Verinin histogram dağılımı sparse hal almaktadır.

Veriyi temizledikten sonra aşağıdaki label verisini elde ettik. Bu sayede temizleme işleminin ilk kısmını tamamladık.

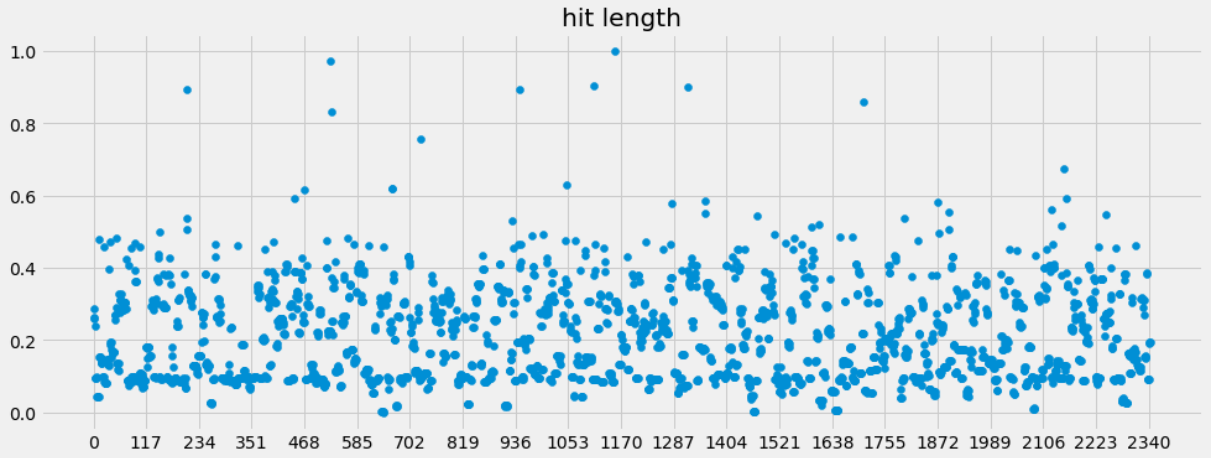


Şekil 6: Temizlenmiş label(‘difficulty’) verisi



Şekil 7: Temizlenmiş label verisinin histogram dağılımı

Her attribute için o attribute'un alınıp alınmayacağını kendimiz karar verdkten sonra kalan verileri inceledik. Aşağıda 'hit_length' attribute için verinin dağılımı görülmektedir.



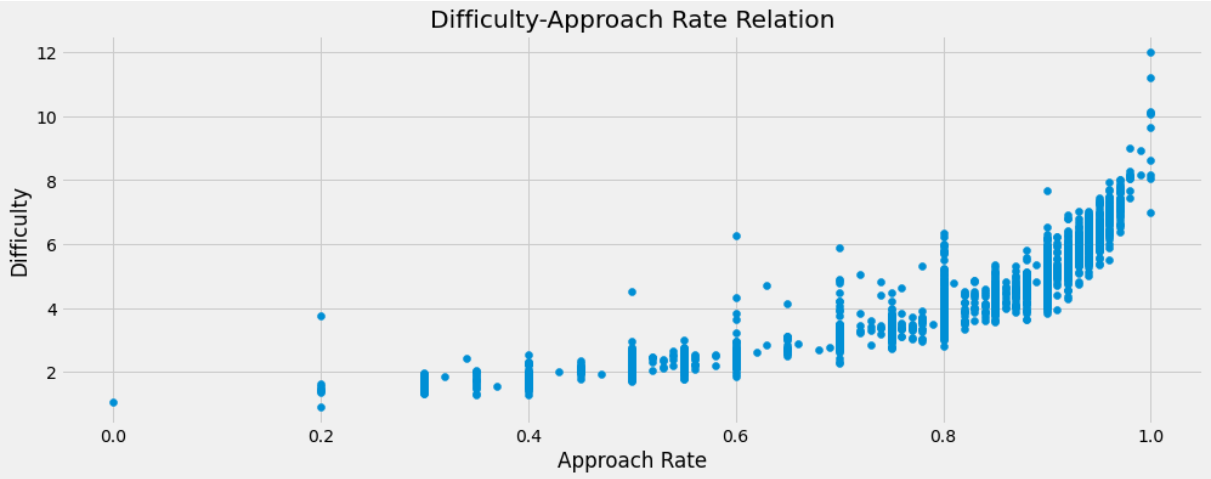
Şekil 8: 'hit_lenght' değeri dağılımı

Her bir sütündaki verilerin nasıl etkili olduğunu anlamak için bu veriler label verisi ile karşılaştırılır.

Hangi Sütün Ne Kadar Etkili ?

Her bir sütündaki değeri karşılaştırmak için ANOVA yani varyans testi uyguladık. Her bir sütünün label verisi ile nasıl bir korelasyon yaptığını inceledik ve veri içindeki grupları label('difficulty') verisini truncate ederek elde ettik. Böylece elimizde 12 tane zorluğa ait incelenecek t testi bulunmaktadır.

Aşağıda Örnek bir sütuna ait korelasyon grafiği görülmektedir. Bu grafiğin model için önemli olup olmadığı F-testi ile incelenebilir. Tüm y lere bir bias veriyorsa F testinin p değeri 0.05 den küçük çıkacaktır.



Şekil 9: Örnek Korelasyon Grafiği

Örnek korrelasyon verisi için ANOVA testi yaparsak; F-testi için aşağıdaki tabloyu elde ederiz. Bu tabloda 13 zorluk derecesi içinden 13-1 den 12 degree of freedom elde edilir. Ve veri seti içerisindeki 2444 veriden veri grubu sayısı yani 13 çıkarılarak 2331 degree of freedom her bir birey için elde edilir. F testi Grupsal varyansın bireysel varyansa bölümü ile bulunur ve o sütünün önemli olup olmadığı bilgisini verir.

	Source	ddof1	ddof2	F	p-unc	np2
0	difficulty	12	2331	26.312953	2.257292e-56	0.119299

Şekil 10: ANOVA F-test

Difficult değerleri arasında p değeri 0.05 den küçük olduğu için anlamlı fark vardır yani bu sütunun veriyi etkilediği söylenebilir. Tüm gruplar kendi içinde etkisini söylemek için t test uygulanır. Bunun için pairwise_tukey test kullanılır ve gruplar kendi içinde incelenir.

Örneğin yukarıdaki veri için aşağıdaki çıktıyı elde ettik.

	A	B	mean(A)	mean(B)	diff	se	T	p-tukey	hedges
0	0.0	1.0	0.052101	0.133065	-0.080965	0.119529	-0.677364	9.999790e-01	-0.676085
1	0.0	2.0	0.052101	0.166126	-0.114026	0.119236	-0.956305	9.992061e-01	-0.955746
2	0.0	3.0	0.052101	0.182149	-0.130048	0.119219	-1.090833	9.971405e-01	-1.090271
3	0.0	4.0	0.052101	0.208130	-0.156029	0.119209	-1.308866	9.853236e-01	-1.308248
4	0.0	5.0	0.052101	0.229933	-0.177832	0.119228	-1.491527	9.581584e-01	-1.490703
5	0.0	6.0	0.052101	0.274551	-0.222450	0.119390	-1.863227	8.166671e-01	-1.860878
6	0.0	7.0	0.052101	0.331408	-0.279307	0.120025	-2.327070	4.971603e-01	-2.317150
7	0.0	8.0	0.052101	0.399095	-0.346994	0.123594	-2.807532	1.991871e-01	-2.727544
8	0.0	9.0	0.052101	0.129412	-0.077311	0.145865	-0.530017	9.999986e-01	-0.000000

Şekil 11: t-testi ilk ikili zorluk('difficulty') grupları

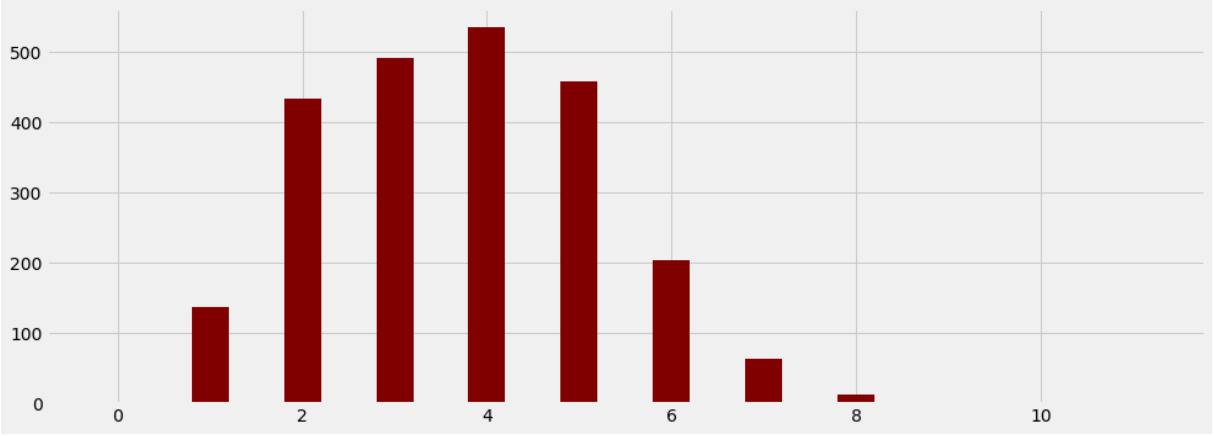
‘0’ zorluğu ve diğer elemanlar arasında fark edilebilen bir farklılık görünmemektedir yani p-değeri istatiksel olarak önemsizdir. Bu da veri eğitimine bir katkı sağlamayacağını göstermektedir.

42	4.0	5.0	0.208130	0.229933	-0.021803	0.007582	-2.875689	1.694475e-01	-0.182927
43	4.0	6.0	0.208130	0.274551	-0.066421	0.009800	-6.777501	1.201905e-09	-0.557131
44	4.0	7.0	0.208130	0.331408	-0.123278	0.015753	-7.825860	0.000000e+00	-1.033790
45	4.0	8.0	0.208130	0.399095	-0.190965	0.033431	-5.712244	9.724127e-07	-1.601221
46	4.0	9.0	0.208130	0.129412	0.078718	0.084372	0.932986	9.993805e-01	0.660026
47	4.0	10.0	0.208130	0.118768	0.089363	0.068954	1.295975	9.865059e-01	0.749276
48	4.0	11.0	0.208130	0.146218	0.061912	0.119209	0.519351	9.999989e-01	0.519106
49	4.0	12.0	0.208130	0.137815	0.070315	0.119209	0.589844	9.999955e-01	0.589565

Şekil 12: t-testi son ikili zorluk(‘difficulty’) grupları

Verileri Uçlardan Kırpılması

Aynı durum 9, 10, 11,12 zorluklu haritalar için de geçerlidir. Bu gruptaki veriler modele katkı sağlamayacaktır bunun sebebinin yetersiz veri olabileceğini düşündük ve veri setindeki harita sayısını zorluk derecelerine göre çizdirdik. Şekil aşağıdaki gibidir ve buradan 9,10,11 ve 12 zorluklu haritaları eğitim setinden atmaya karar verdik çünkü bu veriler eğitim sırasında modeli etkileyebilmektedir Örneğin ardışık iki eğitimden biri yüzde 80 iken sonraki eğitimde bu yüzde 50 ye düşebilmektedir. Bu şekilde outlier lardan kurtulmuş olduk.



Şekil 13: Bazı zorluk verilerin o haritalar hakkında bilgi çıkarmak için yetersiz olduğu görülmektedir

Bu eğitimden sonra aşağıdaki one-way ANOVA metriklerini elde ettik. F değeri arttı bu da grupların etkisi artması anlamına gelmektedir ya da bireylerin etkilerinin azaldığı anlamına gelmektedir.

	Source	ddof1	ddof2	F	p-unc	np2
0	difficulty	7	2328	44.401999	3.006242e-59	0.117785

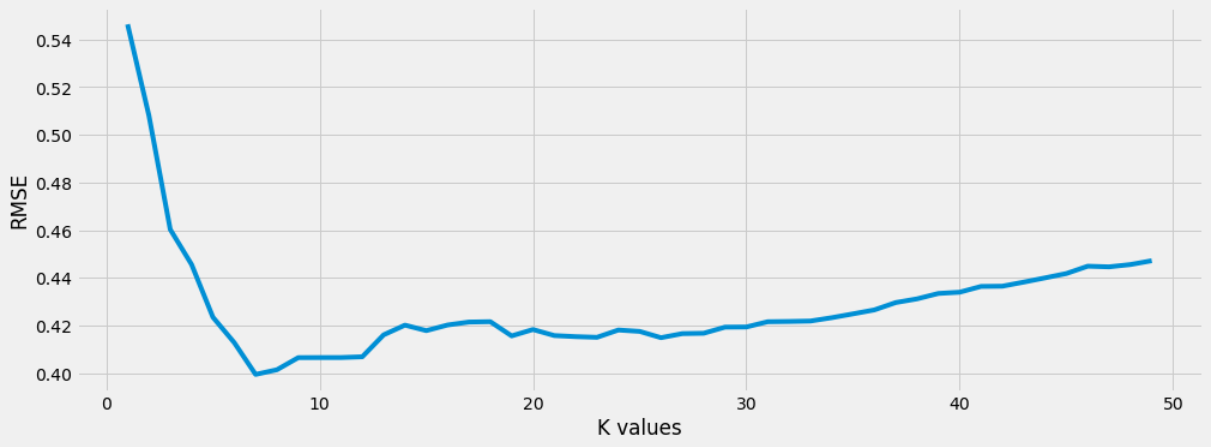
Şekil 14: Temizlenen veriye ait F-testi

Bunu anlamak için ikili Tuckey testini uyguladık ve aşağıdaki sonucu elde ettik.

	A	B	mean(A)	mean(B)	diff	se	T	p-tukey	hedges
0	1.0	2.0	0.133065	0.166126	-0.033061	0.011649	-2.838040	8.631404e-02	-0.277064
1	1.0	3.0	0.133065	0.182149	-0.049084	0.011482	-4.274936	5.258655e-04	-0.411391
2	1.0	4.0	0.133065	0.208130	-0.075065	0.011378	-6.597513	1.441994e-09	-0.629195
3	1.0	5.0	0.133065	0.229933	-0.096867	0.011572	-8.370721	0.000000e+00	-0.811829
4	1.0	6.0	0.133065	0.274551	-0.141486	0.013135	-10.771834	0.000000e+00	-1.184645

Şekil 15: t-testi ikili zorluk('difficulty') grupları

Buradan her ikilinin p değeri 0.05 in altında çıkmaktadır bu da grupların etkisi artmış anlamına gelmektedir.



Şekil 16: K nearest neighbors regresyon algoritması için en iyi k değerinin belirlenmesi

Çeşitli Çalıştırma Örnekleri:

Modeli çalıştırırken on katlı çapraz doğrulama kullanıldı. Hazır modeller ile veri seti denendi ve sırasıyla verilen modeller için şu yüzdeler elde edildi.

Lineer Regresyon Modeli : `0.8723621451708453`

KNN Regresyon Modeli : `0.92701960657592`

Karar Ağacı Regresyon Modeli : `0.9056228712828578`

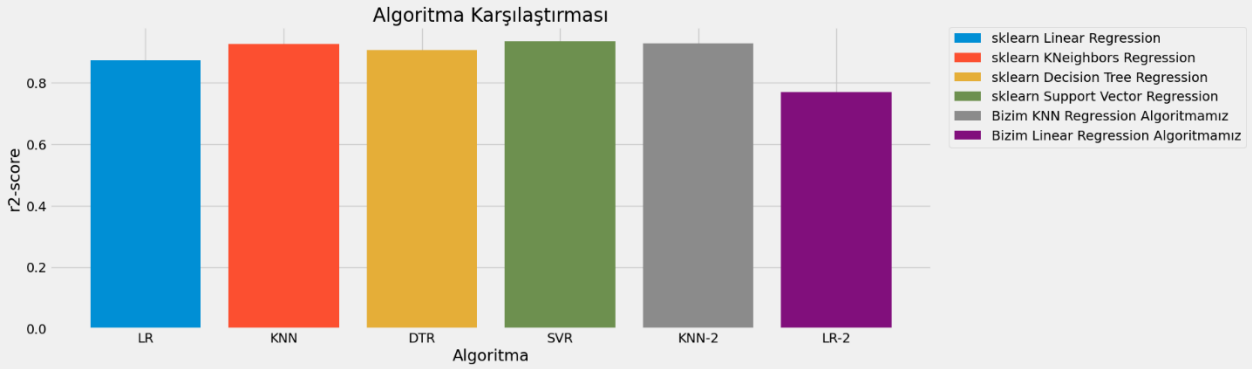
Bayes Ridge Modeli : `0.8723676865758427`

Support Vector Regresyonu : `0.9341645645032205`

Random Forest Regresyonu : `0.9469360794248433`

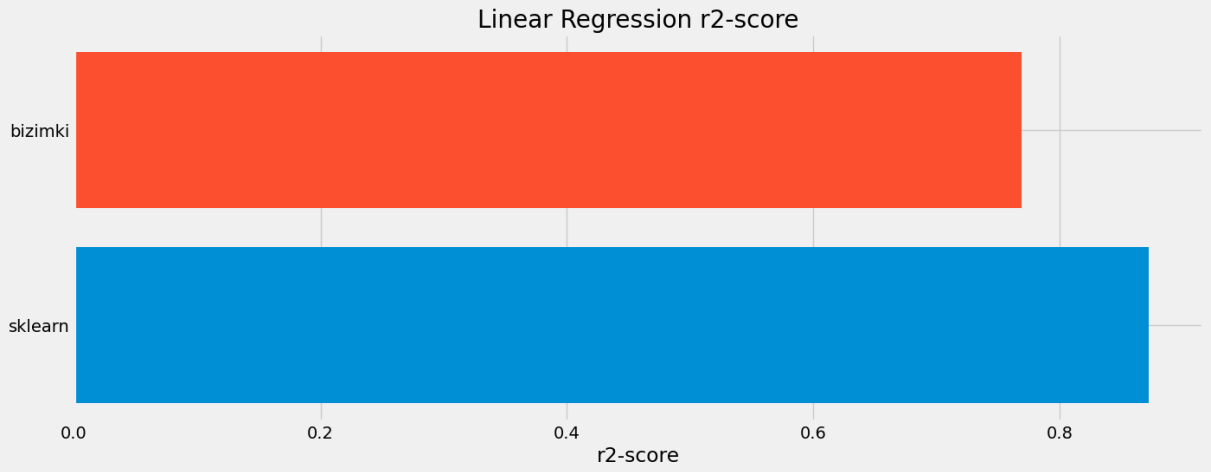
Kendimiz Yukarıdaki İki yaklaşımı kendimiz yazdık ve kendi yazdığımız knn regresyonu için rmse(Root mean square error) değerini `0.06595658503338411` olarak bulduk. Bu da yukarıdaki model gibi 0.93 kesinlik demektir.

Algoritmaların Başarısının Karşılaştırılması:

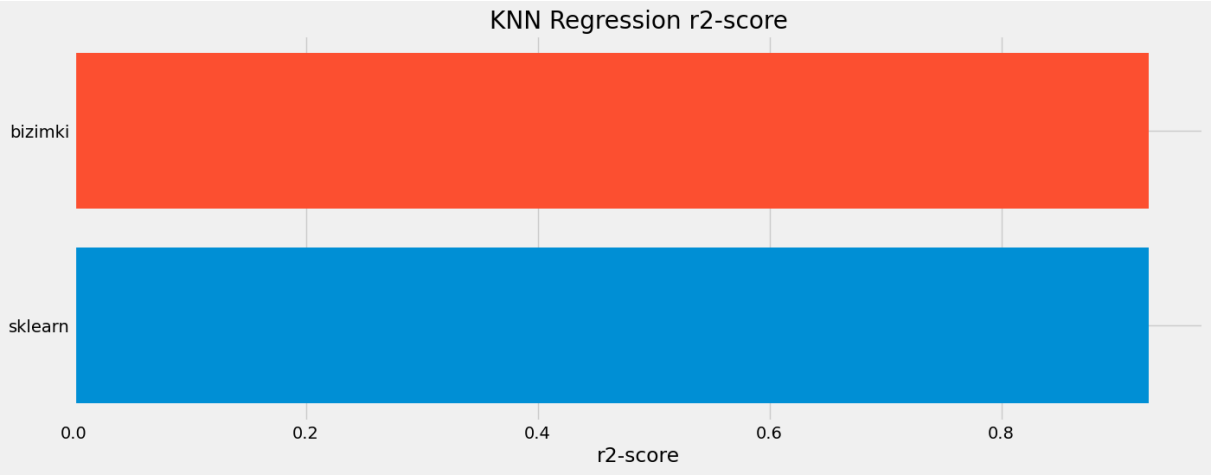


Şekil 17: Algoritma performanslarının karşılaştırılması

Şekil 13'te ilk 4 algoritma sklearn kütüphanesinden gelirken son iki tanesi tarafımızdan gerçeklenmiştir.



Şekil 18: Linear regression algoritmalarının karşılaştırması



Şekil 19: K-nearest neighbors algoritması karşılaştırması

Şekil 18 ve Şekil 19'da kendi algoritmalarımızın sklearn algoritmaları ile karşılaştırmaları görülmektedir. Görüldüğü üzere linear regresyon algoritması geride kalırken k-nearest neighbors algoritması neredeyse birebir aynı sonuç oluşturmuştur.

Sonuç:

Bu projede makine öğrenmesi kullanarak osu! isimli oyunun haritalarının zorluk değerini tahmin etmeyi amaçladık. Veri setini kendimiz elde etmemizin yanında 2 adet regresyon algoritmasını gerçekleyerek bunları hazır sklearn algoritmaları ile kıyasladık. Elde ettiğimiz sonuçlar genel olarak iyi olsa da bu problemin basitliğine yorulabilir. Yine de bu proje bizim için iyi bir deneyimdi ve birçok şey öğrenmiş olduk.