Melih Ocakcı 19011061 Onur Demir 18011078

**Konunun Tanıtımı**

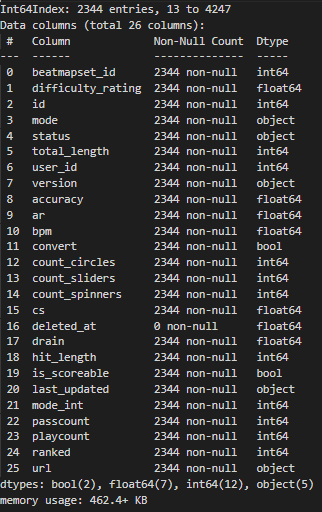
osu! Oyunundaki haritaların verileri internet üzerinden paylaşılmaktadır kişiler bu haritaları indirebilmekte ve bu haritaları oyuna yükleyerek haritayı oynamaktadır. Oyunun haritasını indirdikleri sayfa içerisinde aynı zamanda oyuna ait haritanın bilgisi bulunmaktadır. Bu bilgileri kullanarak o haritaya 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**

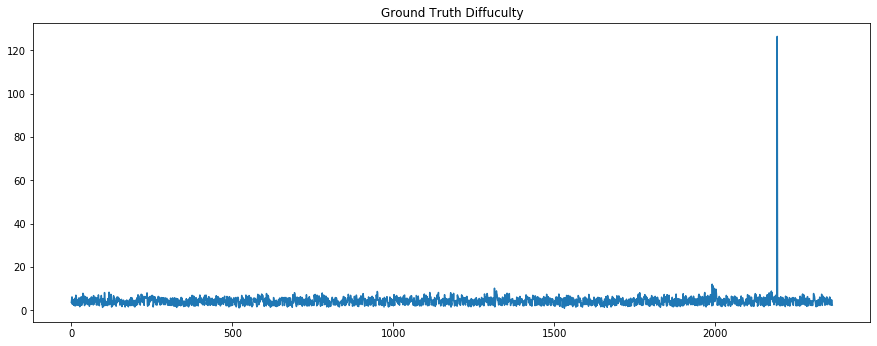
Aşağıdaki şekilde veri setine ait attribute lar verilmiştir. Biz bu attribute lardan 6 tanesini kullanmayı seçtik ve mode olarak sadece osu! nun yuvarlak bazlı haritalarını alacak şekilde filtereledik.



Şekil 0: Ajax çağrısı ile çekilen 25 sütunluk veri

**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.

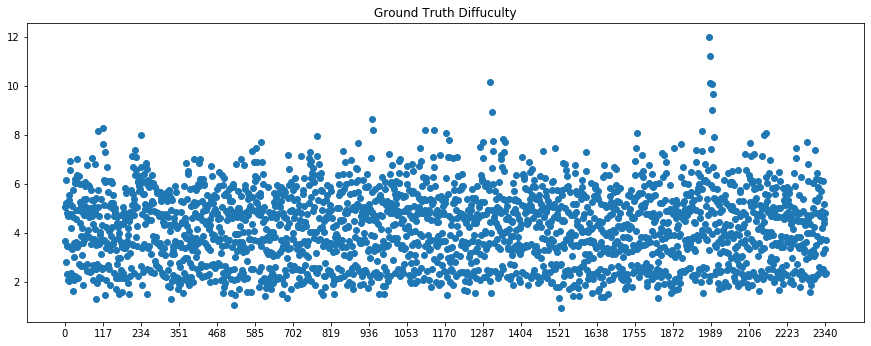


Şekil 1: 125 zorluklu harita model yüzdesini etkilemektedir.

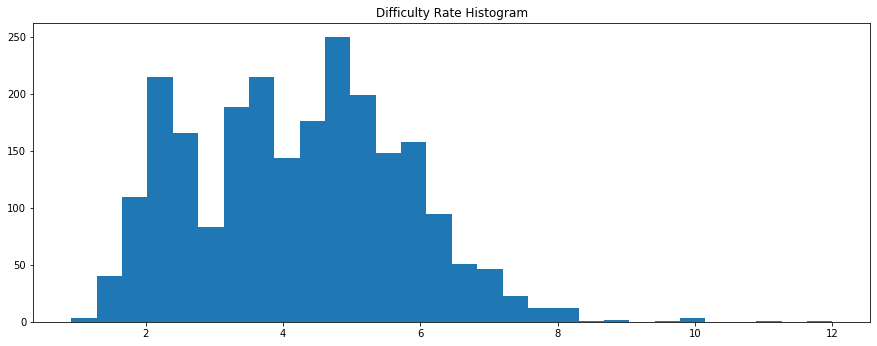


Şekil 2: Verinin histogram dağılımı sparse hal almaktadır.

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

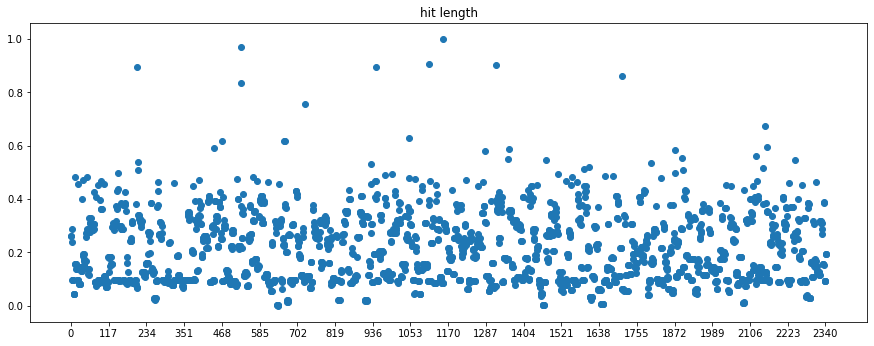


Şekil 3: Temizlenmiş label(‘difficulty’) verisi



Şekil 4: 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.



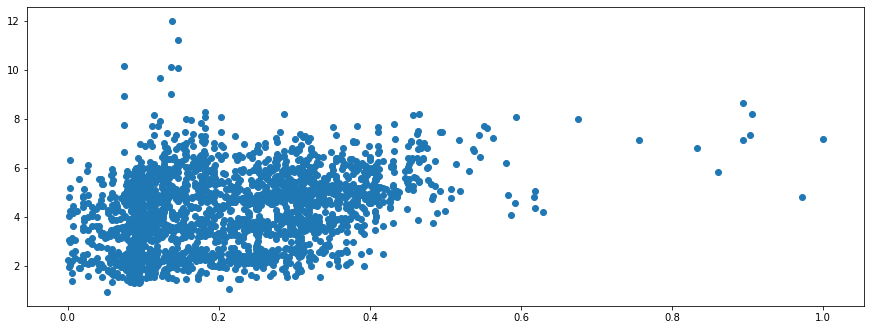
Şekil 5: ‘hit\_lenght’ değeri dağılımı

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 6: Örnek Korelasyon Grafiği

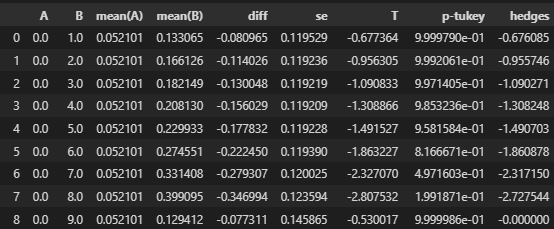
Örnek korrelasyon verisi için ANOVA testi yaparsak; F-testi için aşağıdaki tabloyu elde ederiz. Bu tabloda 13 zorluk derecesi içinden 13-1 den 12 degree of freedom elde edilir. Ve veri seti içerisindeki 2444 veriden veri grubu sayısı yani 13 çıkarılarak 2331 degree of freedom her bir birey için elde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



Şekil 7 : ANOVA F-test

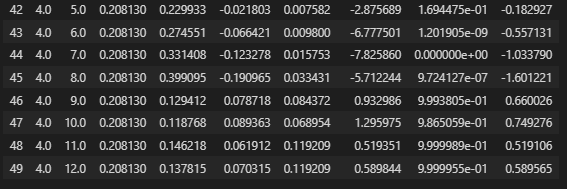
Difficult değerleri arasında p değeri 0.05 den küçük olduğu için anlamlı fark vardır yani bu sütunun veriyi etkilediği söylenebilir. Tüm gruplar kendi içinde etkisini söylemek için t test uygulanır. Bunun için pairwise\_tukey test kullanılır ve gruplar kendi içinde incelenir.

Örneğin yukarıdaki veri için aşağıdaki çıktıyı elde ettik.



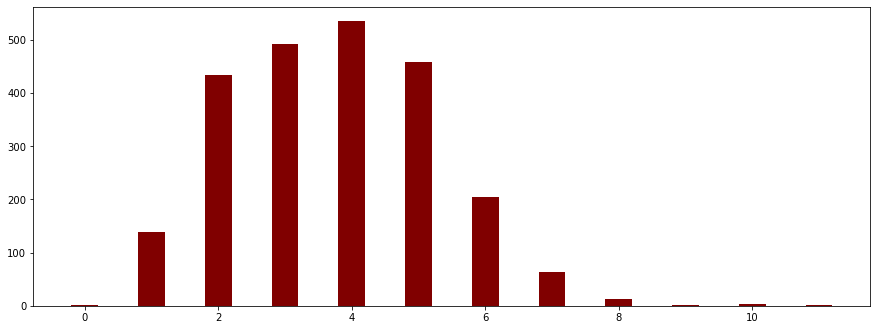
Şekil 8: t-testi ilk ikili zorluk(‘difficulty’) grupları

‘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.



Şekil 9: t-testi son ikili zorluk(‘difficulty’) grupları

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.



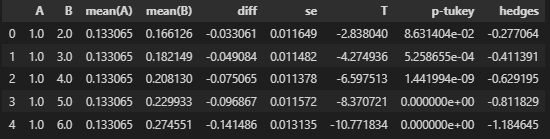
Şekil 10: Bazı zorluk verilerin o haritalar hakkında bilgi çıkarmak için yetersiz olduğu görülmektedir

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.



Şekil 11: Temizlenen veriye ait F-testi

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



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

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

**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.