T.C
DOKUZ EYLÜL ÜNİVERSİTESİ
FEN FAKÜLTESİ

STAJ RAPORU

ALPER ARSLAN

Temmuz 2023
İZMİR

Rapor Değerlendirme

Yaz dönemi zorunlu staj süresi içerisinde Alper ARSLAN tarafından hazırlanmış olan staj raporu tarafımdan okunmuş, kapsamı ve niteliği açısından Staj Raporu olarak kabul edilmiştir.

Prof.Dr.Neslihan DEMİREL Prof.Dr.Güzin ÖZDAĞOĞLU

Teşekkürler

Tüm staj süresince yönlendiriciliği, katkıları ve yardımları ile yanımda duran Prof.Dr. Güzin ÖZDAĞOĞLU'na ve staj bulma süresince bana refans olarak yanımda olan Prof.Dr.Neslihan DEMİREL'e teşekkürlerimi ve saygılarımı iletirim.

Alper ARSLAN

Özet

Yaz dönemi stajı mesleki eğitimi ve becerlilerin gelişmesini amaçlamaktadır. Staj süresince veri setinin toparlanışı, veri setindeki kayıp değerlerin,hatalı veri girişlerin saptanması ve düzeltimesi(Veri Ön İşleme), verinin tanımlayıcı istatistikleri ile birlikte veri setinin yapısına uygun bir biçimde görselleştirilmesi, Veri setinin dağılımının saptanması ve bu duruma bağlı olarak Logaritmik,karekök vb. gibi dönüşüm yöntemleri kullanarak veri setinin dağılımının düzeltilmesi son olarak veri setinde makine öğrenmesi tekniği kullanılarak kümeleme işlemi yapılması ön görülmektedir.

Anahtar Kelimeler: Veri ön işleme, Makine Öğrenmesi, Dağılım, Göreselleştime

Abstract

The summer internship aims at vocational education and the development of skilled workers. The recovery of the data set during the internship, detection of missing values and incorrect data entries in the data set and correction (Data Preprocessing), visualization of the data together with descriptive statistics in a format appropriate to the structure of the data set, Determining the distribution of the data set and depending on this situation, Logarithmic, square root, etc. correction of the distribution of the data set using conversion methods such as finally, it is expected that clustering operation will be performed using machine learning technique in the data set.

Keyword: Data pre-processing, machine learning, Distribution, visualization

İçindekiler

Şekil Listesi

Bölüm 1

1 Giriş

1.1 Times Higher Education (THE) World University Ranking (WUR)

THE WUR sıralaması , dünya üniversitelerinin akademik Sıralaması , QS Dünya Üniversite Sıralaması ve diğerleriyle birlikte genellikle en çok gözlemlenen üniversite sıralamalarından biri olarak kabul edilir . Yayın, konu ve itibara göre üniversitelerin küresel sıralamalarını içerir.

1.2 Veri setinin Tanıtımı

Times Higher Education Dünya Üniversite Sıralamaları, 99 ülkede 1.600'den fazla üniversiteyi içermektedir. THE bir kurumun dört alandaki performansını ölçen 13 performans göstergesine dayanmaktadır.

Performans göstergeleri aşağıdaki gibidir;

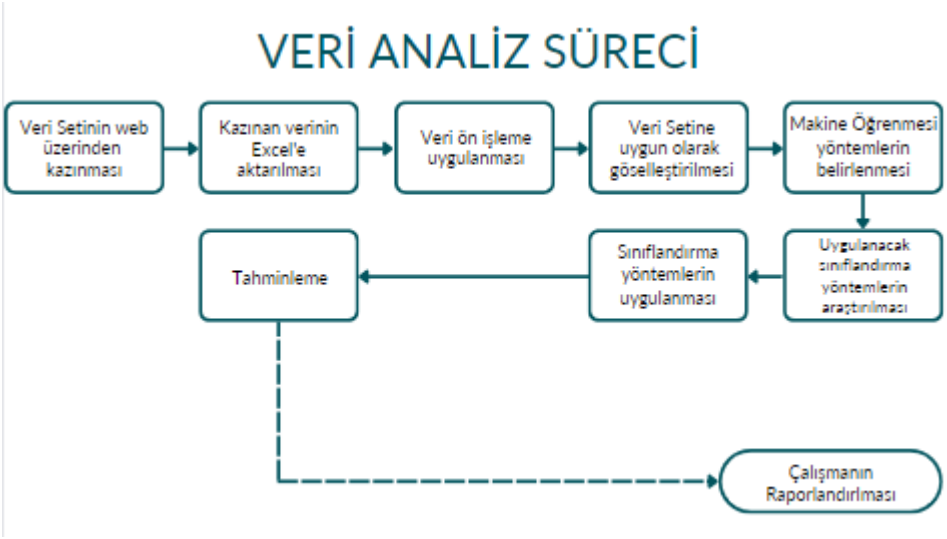
1. **Rank:** Üniversitelerin dünya sıralamasında bulunduğu kategori
    - **1:** 1-50 sıralama sırasında bulunanlar.
    - **2:** 51-100 sıralama sırasında bulunanlar.
    - **3:** 101-150 sıralama sırasında bulunanlar.
    - **4:** 151-200 sıralama sırasında bulunanlar.
    - **5:** 201-250 sıralama sırasında bulunanlar.
    - **6:** 251-300 sıralama sırasında bulunanlar.
    - **7:** 301-350 sıralama sırasında bulunanlar.
    - **8:** 351-400 sıralama sırasında bulunanlar.
    - **9:** 401-500 sıralama sırasında bulunanlar.
    - **10:** 501-600 sıralama sırasında bulunanlar.
    - **11:** 601-800 sıralama sırasında bulunanlar.
    - **12:** 801-1000 sıralama sırasında bulunanlar.
    - **13:** 1001-1200 sıralama sırasında bulunanlar.
    - **14:** 1201-1505 sıralama sırasında bulunanlar.
    - **15:** 1500+ sıralama sırasında bulunanlar.

2. **Name/ Country:** Üniversitenin ismi ve bulunduğu ülke
3. **No. of FTE Student:** üniversitelere kayıtlı öğrenci sayısı
4. **Student perr staff:** Öğretmen başına düşen öğrenci oranı
5. **İnt. Sutudent:** Yabancı öğrenci oranı
6. **Female:** Kadın öğrenci oranı
7. **Male:** Erkek öğrenci oranı
8. **overall:** Genel puan
9. **Teaching:** Öğretim puanı
10. **Research:** Araştırma puanı
11. **Citations:** Atıf puanı
12. **industry income:** Endüstriyel gelir puanı
13. **int. outlook:** Uluslararasılaşma puanı

2 Yöntem ve Teknik

2.1 Veri Analizi süreci

THE sıralama skorları göz önünde alınarak, Dokuz Eylül Üniversitesi'nin dünya sıralama sistemindeki mevcut durumunu incelenmiş, performansını artırmak için Makine Öğrenmesi sınıflandırma yöntemlerini kullanmak ve bu yöntemlerle belirlenen performans artırıcı kriterleri stratejik olarak yorumlamak ve uygulamak üzerine odaklanılmıştır. Üniversitenin sıralamasını artırmak için veriye dayalı ve özgün bir yaklaşım sunmak, performans artırıcı kriterleri belirlemek ve bu kriterleri stratejik bir şekilde yorumlayarak uygulanabilir eylemler önermektir. Bu şekilde, Dokuz Eylül Üniversitesi'nin uluslararası düzeyde rekabet edebilirliği ve itibarı artırılabilir ve sıralama sistemlerinde üst seviyelere çıkartılması hedeflenmektedir.



Şekil 1: Süreç Şeması

2.2 Naive Bayes Sınıflandırma

Naive Bayes sınıflandırıcısı veri madenciliği, makine öğrenmesi, duygu analizi gibi alanında yaygın kullanıma sahip denetimli öğrenme sınıfında bir algoritmadır. Olasılık hesaplamalarını kullanarak en yüksek olasılığa sahip kararı seçmeyi amaçlar.Bu sınıflandırıcı bir takım basit hesaplamalar yaparak olasılık tabanlı olarak bir olayın gerçekleşme ihtimalini hesaplar.(Şahinaslan vd.,2022)

Bayes karar teoremi eşitliği aşağıda yer almaktadır.

$$P(S_i|X) = \frac{P(X|S_i)P(S_i)}{P(X)}$$

$P(S_i|X)$: Seçilen X değerinin i'inci sıra grubunda olma olasılığı.
$P(S_i)$: i sıra grubuna ait ilk olasılığı.
$P(X)$: Seçilen bir örneğin X olmasının ilk olasılığı.
$P(X)$: i sıra grubuna ait bir örneğin X olma olasılığı.

2.3 Decision Tree Sınıflandırma

Karar ağaçları, diğer sınıflandırma yöntemlerine göre kolay yorumlanabilmeleri, daha düşük maliyetlerle gerçekleştirilebilmeleri, veri tabanları ile entegrasyon kolaylığı ve güvenilirlik düzeylerinin iyi olması nedeniyle sıklıkla kullanılan bir sınıflandırma ve regresyon tekniğidir (Chein and Chen, 2008). Ayrıca karar ağaçlarında karar kuralı olarak gösterilen yapraklar bu alanda çalışan kişiler tarafından kolaylıkla yorumlanabilmekte ve bu yöntem yüksek boyutlu verilerde etkin olarak kullanılmaktadır (Aytekin et al., 2018). Karar ağaçları, değinilen üstün özelliklerine karşın, birden fazla öznitelik içeren çıktıları olanaklı kılmamaları, kısmen değişken sonuçlar üretmeleri, test verisindeki küçük değişikliklere karşı bile duyarlı olmaları, nümerik veri setleri için karmaşık bir ağaç yapısı oluşturmaları gibi problemler ile karşı karşıya kalmaktadır.(Zhao ve Zhang, 2008)

2.4 Random Forest Sıralaması

Random Forest algoritması, eğitim verisindeki örneklerin rastgele olarak seçilmesi ile oluşturulan budanmamış sınıflandırma ve regresyon ağaçlarından oluşan bir modeldir. Bu modelde, sınıflandırıcıların genelleştirme hatası, tüm ağaçların bireysel gücüne ve bu ağaçlar arasındaki bağıntıya dayalıdır (Breiman,2001).

2.5 KNN Sınıflandırma

K-NN algoritması, en temel örnek tabanlı öğrenme algoritmaları arasındadır (Mitchell, 1997). K-NN algoritması basit bir yapıya sahiptir ve az sayıda parametre gerektirmektedir (Coomans ve Massart 1982). K-NN algoritması, büyük eğitim setlerinin varlığında, oldukça etkin sonuçlar verebilmektedir. K-NN algoritması, ilgisiz özniteliklerin varlığında da sınıflandırma modeli oluşturabilmektedir (Aha vd. 1991). K-NN algoritması basit yapısına karşın, yüksek bir hesaplama maliyetine sahiptir. Sınıf etiketi belirlenmek istenen örneğin, veri setinde yer alan örnekler ile arasındaki uzaklığın belirlenmesi, özellikle büyük eğitim veri setleri için oldukça maliyetli olabilmektedir. Bu maliyeti ortadan kaldırmak için, K-NN algoritması temel bileşenler analizi gibi boyut azaltma yöntemleri ile ya da arama ağaçları gibi daha güçlü veri yapıları ile birlikte kullanılabilmektedir (Coomans ve Massart 1982).

Öklid uzaklığı, sınıflandırma ve kümeleme algoritmalarında en sık kullanılan uzaklık ölçütüdür:

$$\sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$$

Öklid uzaklığı, K-ortalama kümeleme algoritması, temel K-NN algoritması gibi sınıflandırma ve kümeleme algoritmalarında yakınlığın ölçümlenmesi için kullanılan temel uzaklık ölçütüdür.

2.6 Lojistik Regresyon Sınıflandırma

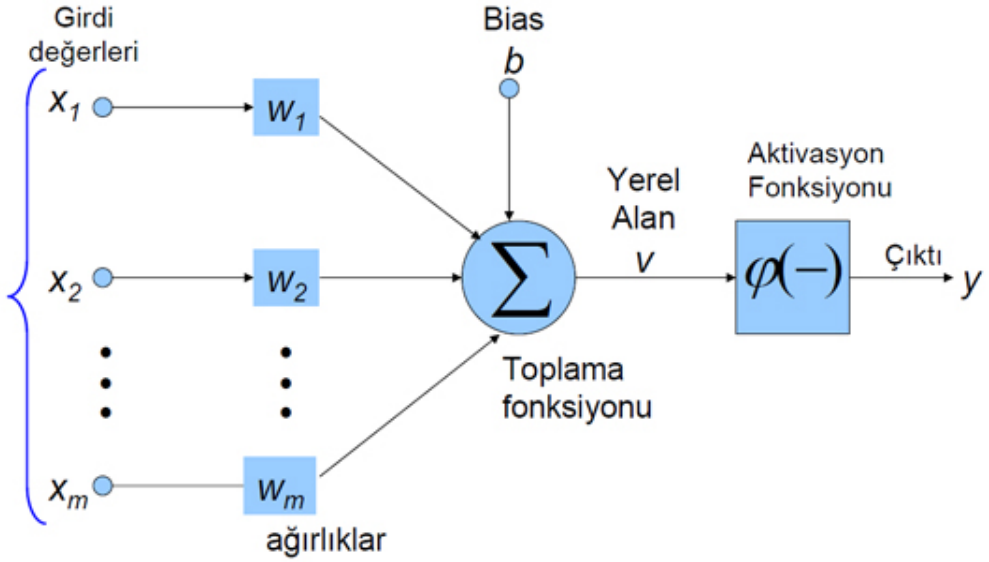
Lojistik regresyon analizinde logit dönüşümün uygulandığı bağımlı (yordayıcı) değişkenin yapısı analizin sınıflandırılmasında önemli bir yere sahiptir.

Çok değişkenli istatistiksel verilerin sınıflandırılmasında kullanılan yöntemlerden biri olan lojistik regresyon analizinde verilerin yapısındaki grup sayısı bilinmekte ve bu verilerden hareketle bir ayrımsama modeli oluşturulmaktadır (Ulupınar, 2007).

Lojistik regresyon yönteminde bağımlı değişkenin sürekli olması gibi bir varsayım yoktur, özellikle bağımlı değişkenin iki veya daha çok kalitatif değer aldığı durumlarda kullanılır (Ulupınar, 2007).

2.7 Yapay Sinir Ağları

Yapay Sinir Ağları (Artificial Neural Networks) veya kısaca Sinir Ağları (YSA) insan beyninden esinlenerek geliştirilmiş, ağırlıklı bağlantılar aracılığı ile birbirine bağlanan işlem elemanlarından oluşan paralel ve dağıtılmış bilgi işleme yapılarıdır.



Şekil 2: Yapay Sinir Ağ yapısı

Yapay sinir ağlarının temel işlevleri şu şekilde belirtilebilir :

- Öngörü (Prediction) veya tahminleme : İleriki satışlar, hava tahminleri, at yarışları, çevresel risk, ...
- Sınıflandırma (Classification) ve Kümeleme (Clustering) : Müşteri profilleri, tıbbi teşhis, ses ve şekil tanıma, hücre tipleri ...
- Kontrol (Control) : Erken uyarı için uçaklarda ses ve titreşim düzeyleri, ...

Ayrıca, Veri Birleştirme (Data Association), Kavramsallaştırma (Data Conceptualization) ve Filtreleme (Data Filtering) için de kullanılabilir. (Uğur ve Kınacı, 2006)

2.8 Gradient Boost Machine (GBM) Sınıflandırma

Gradient Boosting Machine (GBM), 2001 yılında Friedman tarafından önerilen bir yinelemeli birleştirme prosedürüdür, ve güçlü bir topluluk oluşturmak için birden çok zayıf öğreniciyi birleştirir. GBM, genellikle denetimli görevlerde (sınıflandırma veya regresyon) kullanılır.

Bu yöntemde, zayıf öğreniciler (genellikle karar ağaçları) önceki öğrenicilerin hatalarını düzeltmeye çalışarak ardışık olarak eğitilir. İlk öğrenici, veri kümesi

üzerinde eğitilir ve ardından hataları hesaplanır. Sonraki öğrenici, önceki öğrenicinin hatalarını düzeltmeye odaklanarak, kalan hatayı azaltmak için eğitilir. Bu süreç, belirli bir durma kriterine ulaşılana veya belirli bir maksimum öğrenici sayısına ulaşılıncaya kadar tekrarlanır.

Sonuç olarak Gradient Boosting Machine, yinelemeli bir birleştirme prosedürü, sınıflandırma veya regresyon gibi denetimli görevlerde kullanılarak birden çok zayıf öğreniciyi güçlü bir topluluk oluşturmak için bir araya getirir.

2.9 Cat Boost Sınıflandırma

CatBoost, kategorik boosting fikri üzerine inşa edilmiştir. Algoritma bir bölünme seçilmeden önce ağaç oluşturma işlemini gerçekleştirirken, tüm kategorik özellikleri sayısal hale dönüştürür. Bu dönüşüm, yalnızca kategorik özellikler ve kategorik ve sayısal özellikler birlikte kombinasyonları olmak üzere her iki kombinasyonda da birkaç istatistik uygulanarak yapılır (Alshari vd., 2021).

CatBoost'un bu özelliği sayesinde kategorik verilerle daha iyi başa çıkabilmesini sağlar. Özellikle kategorik değişkenler çok sayıda kategoriye sahip olduğunda veya kategorik değişkenler ile sayısal değişkenler arasında güçlü bir etkileşim olduğunda önemlidir. CatBoost'un kategorik boosting yaklaşımı, onu diğer boosting algoritmalarından ayırır. Bu yaklaşım, CatBoost'un daha yüksek bir doğruluk ve daha iyi bir genelleştirici performans elde etmesini sağlar.

CatBoost, kategorik verilerle çalışan bir makine öğrenmesi algoritması olduğundan, kategorik verileri daha iyi anlayabilir ve daha iyi tahminler yapabilir.

3 Uygulama

3.1 Veri Ön İşleme

Veri setimiz, THE sistemi tarafından skorlandırılarak sıralama yapılmıştır. Veri seti üzerinde hatalı veri girişi tespit edilmemiş, ekstra bir düzenleme ihtiyacı ortaya çıkmamıştır. Ayrıca, kayıp değer araştırması yapılmış ve bu kayıp değerlerin tek bir değişkende olduğu belirlenmiştir. Kayıp değerler, ilgili değişkenin yapısına uygun bir şekilde ortalamaları kullanılarak doldurulmuştur.

```
Rank                      0
Name/ Country             0
No. Of FTE Students       0
Students per staff        0
int. Students             0
Female                   85
Male                     85
Overall                   0
Teaching                  0
Research                  0
Citations                 0
industry income           0
int.outlook               0
dtype: int64
```

Şekil 3: Kayıp Değer Tespiti

Veri ön işleme sürecinde, veri setimizde bulunan kayıp değerlerin tek bir değişkende sınırlı olması, analiz ve sonuçların güvenilirliğini artırmıştır. Ayrıca, uygun bir doldurma yöntemi kullanılarak eksik verilerin giderilmesi, veri analizini ve yorumlamayı olumsuz etkileyecek boşlukların önüne geçmiştir.

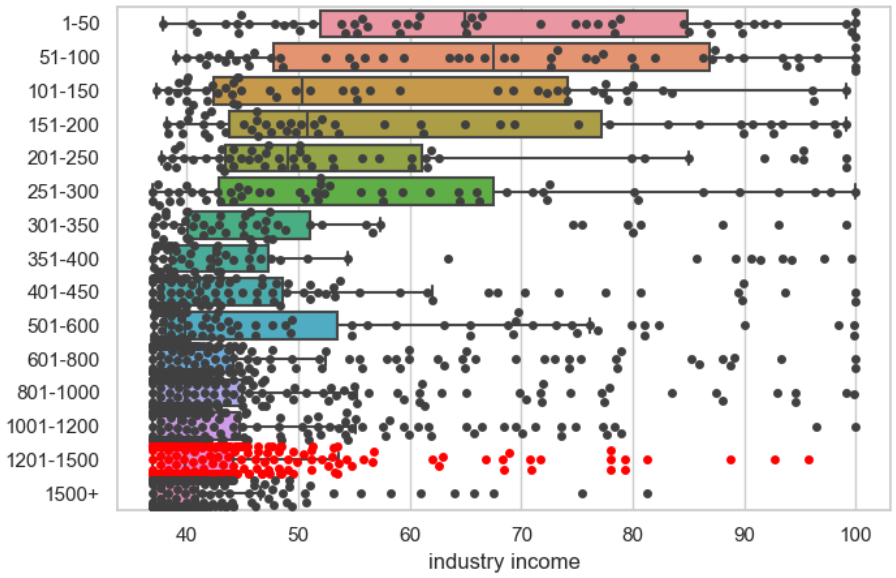
3.2 Veri Görselleştirme ve Keşifsel analiz

Veri setinin ön işleme adımları tamamlandıktan sonra Keşifsel analizleri yapılmıştır. Değişkenlerin ortalamarı, medyanları, standart sapmaları vb. değerleri hesaplanmış ve tablo halinde Şekil 4'de gösterilmiştir.

	No. Of FTE Students	Students per staff	Female	Male	Teaching	Research	Citations	industry income	int.outlook
count	1799.000000	1799.000000	1714.000000	1714.000000	1799.000000	1799.000000	1799.000000	1799.000000	1799.000000
mean	22505.192885	18.508338	50.914236	49.085764	27.018010	23.016898	48.495887	47.104558	46.880378
std	26808.634140	11.125834	12.409259	12.409259	13.282243	16.763819	27.967185	15.093682	22.582401
min	489.000000	0.900000	2.000000	0.000000	11.600000	7.400000	0.800000	36.900000	14.100000
25%	9515.500000	12.400000	44.000000	42.000000	18.000000	11.300000	23.100000	37.800000	27.900000
50%	17001.000000	16.200000	53.000000	47.000000	22.700000	17.000000	47.200000	40.500000	42.100000
75%	28578.000000	21.650000	58.000000	56.000000	31.850000	28.900000	72.350000	48.300000	62.100000
max	460632.000000	232.200000	100.000000	98.000000	94.800000	99.700000	100.000000	100.000000	99.700000

Şekil 4: Keşifsel Analiz Tablosu

Keşifsel analizin ardından üniversitelerin grupları ve rastgele ele alınan bir değişkenin görselleştirilmesi yapılmıştır. Bu görselleştirmede, Dokuz Eylül Üniversitesi'nin dahil olduğu gruptaki üniversiteler farklı bir renkle belirtilmiştir ve veri setinin analizine katkı sağlanmıştır.



Şekil 5: Veri Seti Ön İzleme Grafiği

3.3 Sınıflandırma Yöntemleri

3.3.1 Train Test Split

Train-test-split uygulama amacımız veri setini Train(eğitim) ve Test(değerlendirme) olacak şekilde ikiye ayırmaktır. Bu yöntem sayesinde eğitilen modelin gerçek dünya verilerine genelleme yapabilme özelliği kazandırmaktır.

```
# Traning - Test

y=df["Rank"]
X=df.drop(["Rank","Overall","Male"],axis=1)

x_train,x_test,y_train,y_test=train_test_split(X,y,test_size=0.20,random_state=42,stratify=y)
```

Şekil 6: Train Test Split Uygulama Kodu

Şekil 6' da kod dizisinde bağımlı değişken olan Rank y değişkenine atanmıştır, ardından bağımsız değişkenlere atanması gerekenler belirlenip atanmayacaklar dışarıda bırakılmış halde x olarak tanımlanmıştır. Atama işlemleri tamamlandıktan sonra X train, X test, Y train ve Y test olacak şekilde veri seti ikiye bölünmüştür bu işlemi uygularken test setinin, veri setinin %20' si olacak şekilde ayırmaya dikkat edilmiştir.

3.3.2 Naive Bayes

Veri seti train edildikten sonra Navie bayes sınıflandırma yöntemi ile veri seti eğitilmiş ve performansını test etmek amacı ile Classication report kütüphanesi kullanılarak accuracy,recall ve precision performans parametreleri ölçülmüştür.

	precision	recall	f1-score	support
1	1.00	0.70	0.82	10
2	0.62	0.80	0.70	10
3	0.43	0.60	0.50	10
4	0.18	0.20	0.19	10
5	0.57	0.40	0.47	10
6	0.25	0.20	0.22	10
7	0.00	0.00	0.00	10
8	0.00	0.00	0.00	10
9	0.27	0.48	0.34	21
10	0.17	0.16	0.16	19
11	0.45	0.60	0.52	40
12	0.56	0.25	0.34	40
13	0.55	0.68	0.61	41
14	0.85	0.85	0.85	61
15	0.96	0.93	0.95	58
accuracy	0.58	360		
macro avg	0.46	0.46	0.45	360
weighted avg	0.58	0.58	0.57	360

Şekil 7: Naive Bayes Performans Testi

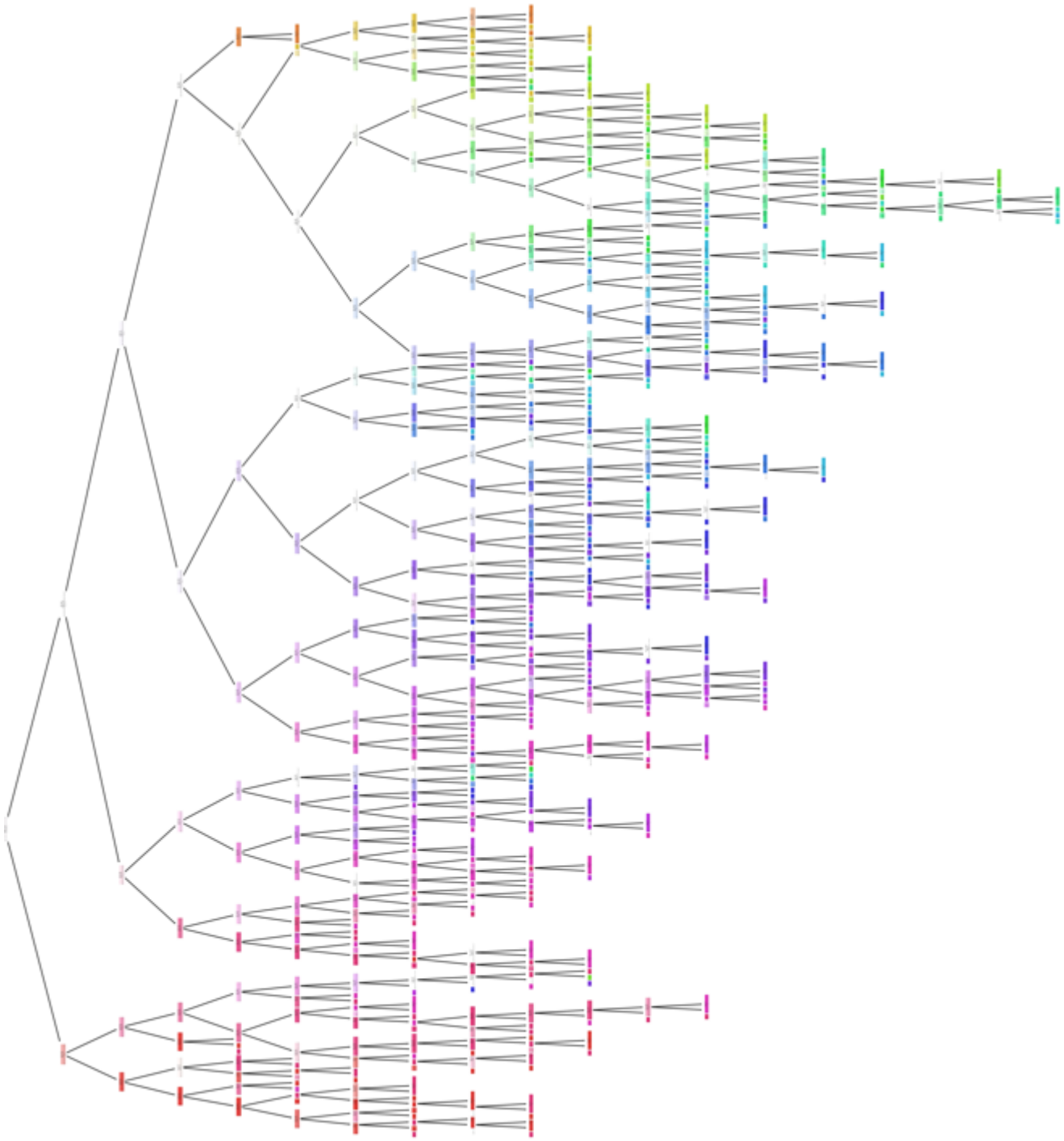
Precision parametresi, her sınıf için pozitif tahminlediğimiz değerlerin gerçekten kaç adetinin pozitif olduğunu göstermektedir.

Recall parametresi pozitif olarak tahmin etmemiz gereken işlemlerin ne kadarını pozitif olarak tahmin ettiğimizi gösteren bir metriktir. Accuracy parametresi, modelin başarısını tahminleyen bir metriktir.

Naive Bayes sonuçlarına dayanarak, test edilen modelin accuracy,precision ve recall performans parametreleri, her sınıf için farklı performans değerleri elde ettiği gözlemlenmiş ve genel olarak modelin ortalama bir performans sergilediği görülmüştür.

3.3.3 Decision Tree

Veri seti train edildikten sonra Karar ağacı (Decision Tree) sınıflandırma yöntemi kullanılarak model oluşturulmasına karar verilmiştir.Oluşturulan model ile veri seti eğitilmiş ve incelenmiştir.



Şekil 8: Karar Ağacı Grafiği

İnceleme neticesinde karar ağacı modelinin 629 adet düğüm, 315 yapraktan oluştuğu görülmüştür. Modelin performansını test etmek amacı ile Classication report kütüphanesi kullanılarak accuracy,recall ve precision performans parametreleri ölçülmüştür.

	precision	recall	f1-score	support
1	0.89	0.80	0.84	10
2	0.67	0.60	0.63	10
3	0.62	1.00	0.77	10
4	0.50	0.20	0.29	10
5	0.43	0.30	0.35	10
6	0.38	0.60	0.46	10
7	0.50	0.60	0.55	10
8	0.38	0.30	0.33	10
9	0.36	0.38	0.37	21
10	0.23	0.26	0.24	19
11	0.74	0.57	0.65	40
12	0.68	0.68	0.68	40
13	0.70	0.78	0.74	41
14	0.71	0.69	0.70	61
15	0.80	0.81	0.80	58
accuracy			0.63	360
macro avg	0.57	0.57	0.56	360
weighted avg	0.64	0.63	0.63	360

Şekil 9: Karar Ağacı Performans Tablosu

Modelin Classification report tablosuna bakıldığında, performans kriterleri modelin ortalama bir performans sunduğunu açıklamıştır. Performans arttırmak için modeli geliştirebilir ya da düşük oran veren sınıflar üzerinde çalışılabilir

3.3.4 Random Forest

Random Forest sınıflandırma yöntemi kullanılarak bir model oluşturulmuş ve bu modeli veri setimizi eğitmek üzere kullandık. Random Forest ile oluşturulan modelin performans değerine öncelikle accuracy metriği ile ölçüldü.

Random Forest'ın accuracy performans değerine bakıldığında daha iyi bir performans elde etmek mümkün olunabileceğini düşünerek modelin belirli metrikleri (n estimators:1000, max features: 5 ve min samples split:2) olacak şekilde

```
accuracy_score(y_pred_3,y_test)
```

```
0.7
```

Şekil 10: Random Forest Accuracy Skoru

ayarlanmıştır.

Düzeltilen metrikler;

- ***n* estimators :** Topluluk içerisindeki ağaç sayısını belirtir.
- **max features:** Maksimum özellik sayısı.
- **min samples split:** Bir düğümün bölünebilmesi için en az kaç örnek içermesi gerektiğini belirtir.

Belirlenen metrikler ile yeni bir model oluşturuldu ve performans testi yapıldı.

	precision	recall	f1-score	support
1	1.00	0.80	0.89	10
2	0.75	0.90	0.82	10
3	0.64	0.90	0.75	10
4	1.00	0.20	0.33	10
5	0.47	0.70	0.56	10
6	0.50	0.40	0.44	10
7	0.50	0.50	0.50	10
8	0.67	0.40	0.50	10
9	0.54	0.62	0.58	21
10	0.50	0.53	0.51	19
11	0.74	0.65	0.69	40
12	0.74	0.72	0.73	40
13	0.77	0.90	0.83	41
14	0.83	0.82	0.83	61
15	0.86	0.88	0.87	58
accuracy			0.73	360
macro avg	0.70	0.66	0.66	360
weighted avg	0.74	0.73	0.73	360

Şekil 11: Random Forest Performans Tablosu

Performans tablosuna bakıldığında düzeltilmiş modelin ilkel modele göre daha performanslı olduğu görülmüştür. Düzeltilmiş model performası 0.73 olduğundan ortama üstü bir performans göstermektedir.

3.3.5 K-NN

K-NN (*k*- Nearest Neighbors) sınıflandırma yöntemi kullanılarak bir model oluşturulmuştur. K-NN modeli veri setini eğitere her örneğin en yakın komşusuna bakarak sınıflandırması yapar. Sınıflandırma sonunda modelin başarısını ve performansını gözlemlemek için classification score' una bakılmıştır.

	precision	recall	f1-score	support
1	0.00	0.00	0.00	10
2	0.14	0.10	0.12	10
3	0.20	0.10	0.13	10
4	0.09	0.10	0.10	10
5	0.00	0.00	0.00	10
6	0.00	0.00	0.00	10
7	0.00	0.00	0.00	10
8	0.00	0.00	0.00	10
9	0.12	0.14	0.13	21
10	0.04	0.05	0.05	19
11	0.19	0.20	0.20	40
12	0.11	0.15	0.12	40
13	0.21	0.22	0.21	41
14	0.29	0.26	0.27	61
15	0.32	0.34	0.33	58
accuracy			0.18	360
macro avg	0.11	0.11	0.11	360
weighted avg	0.18	0.18	0.18	360

Şekil 12: K-NN Performans Tablosu

Performans tablosu değerlendirildiğinde genel doğruluk accuracy değeri 0.18 çıktığı gözlemlenmiştir. Bu değerlere dayanarak modelin performansının düşük olduğu söylenmektedir. Performansının fazla düşük olması sebebi ile modeli düzeltme yapılması taktirde değişikliğin minimum olacağı ön görüldüğünden model üzerinde düzeltmeler yapılmamıştır.

3.3.6 Lojistik Regresyon Sınıflandırma

Lojistik Regresyon sınıflandırma yöntemi kullanılarak bir model oluşturuldu.Sınıflandırma sonucunda modelin başarısını ve performansını değerlendirmek için sınıflandırma skoruna bakılmıştır.

Class	Precision	Recall	F1-Score	Support
1	0.75	0.90	0.82	10
2	0.21	0.30	0.25	10
3	0.57	0.40	0.47	10
4	0.00	0.00	0.00	10
5	0.80	0.40	0.53	10
6	0.50	0.10	0.17	10
7	0.00	0.00	0.00	10
8	0.00	0.00	0.00	10
9	0.00	0.00	0.00	21
10	0.50	0.05	0.10	19
11	0.19	0.50	0.28	40
12	0.18	0.20	0.19	40
13	0.18	0.07	0.10	41
14	0.47	0.70	0.56	61
15	0.76	0.78	0.77	58
Accuracy		0.39		360
Macro Avg	0.34	0.29	0.28	360
Weighted Avg	0.37	0.39	0.35	360

Şekil 13: Lojistik Regresyon Performans Tablosu

Performans tablosu değerlendirildiğinde genel doğruluk accuracy değeri 0.39 çıktığı gözlemlenmiştir. Bu değerlere dayanarak modelin performansının ortalamanın altında olduğu söylenmektedir. Performansının 0.60 olması sebebi ile modeli düzeltme yapılması taktirde değişikliğin minimum olacağı ön görüldüğünden model üzerinde düzeltmeler yapılmamıştır.

3.3.7 Yapay Sinir Ağları

Yapay Sinir Ağı sınıflandırma yöntemi kullanılarak bir model oluşturuldu.Sınıflandırma sonucunda modelin başarısını ve performansını değerlendirmek için sınıflandırma skoruna bakılmıştır.

Class	Precision	Recall	F1-Score	Support
1	0.00	0.00	0.00	10
2	0.00	0.00	0.00	10
3	0.00	0.00	0.00	10
4	0.00	0.00	0.00	10
5	0.00	0.00	0.00	10
6	0.00	0.00	0.00	10
7	0.00	0.00	0.00	10
8	0.00	0.00	0.00	10
9	0.00	0.00	0.00	21
10	0.00	0.00	0.00	19
11	0.00	0.00	0.00	40
12	0.00	0.00	0.00	40
13	0.00	0.00	0.00	41
14	0.17	1.00	0.29	61
15	0.00	0.00	0.00	58
Accuracy		0.17		360
Macro Avg	0.01	0.07	0.02	360
Weighted Avg	0.03	0.17	0.05	360

Şekil 14: Yapay Sinir Ağı Performans Tablosu

Performans tablosu değerlendirildiğinde genel doğruluk accuracy değeri 0.17 çıktığı gözlemlenmiştir. Bu değerlere dayanarak modelin performansının ortalamanın altında olduğu söylenmektedir. Performansının 0.60 olması sebebi ile modeli düzeltme yapılması taktirde değişikliğin minimum olacağı ön görüldüğünden model üzerinde düzeltmeler yapılmamıştır.

3.3.8 Gradient Boost Machine Sınıflandırma

Gradient Boost Machine sınıflandırma yöntemi kullanılarak bir model oluşturuldu.Sınıflandırma sonucunda modelin başarısını ve performansını değerlendirmek için sınıflandırma skoruna bakılmıştır.

Class	Precision	Recall	F1-Score	Support
1	1.00	0.80	0.89	10
2	0.82	0.90	0.86	10
3	0.62	0.80	0.70	10
4	0.44	0.40	0.42	10
5	0.62	0.50	0.56	10
6	0.30	0.30	0.30	10
7	0.25	0.20	0.22	10
8	0.29	0.20	0.24	10
9	0.41	0.52	0.46	21
10	0.26	0.26	0.26	19
11	0.61	0.50	0.55	40
12	0.59	0.65	0.62	40
13	0.73	0.78	0.75	41
14	0.83	0.85	0.84	61
15	0.89	0.86	0.88	58
Accuracy		0.66		360
Macro Avg	0.58	0.57	0.57	360
Weighted Avg	0.66	0.66	0.66	360

Şekil 15: GBM İlkel Performans Tablosu

Performans tablosu değerlendirildiğinde genel doğruluk accuracy değeri 0.66 çıktığı gözlemlenmiştir. Bu değerlere dayanarak modelin performansının ortalama olduğu söylenmektedir. Performansının 0.66 olması sebebi ile modeli düzeltme yapılması taktirde değişikliğin olacağı ön görüldüğünden model üzerinde düzeltmeler yapılmıştır. Modelin parametreleri, learning rate : 0.1, max depth : 5 ve n estimators : 500 olarak ayarlanmış ve tekrardan model kurulmuştur.

Class	Precision	Recall	F1-Score	Support
1	1.00	0.80	0.89	10
2	0.82	0.90	0.86	10
3	0.77	1.00	0.87	10
4	0.75	0.30	0.43	10
5	0.64	0.70	0.67	10
6	0.36	0.40	0.38	10
7	0.30	0.30	0.30	10
8	0.44	0.40	0.42	10
9	0.52	0.57	0.55	21
10	0.28	0.26	0.27	19
11	0.63	0.60	0.62	40
12	0.62	0.62	0.62	40
13	0.70	0.80	0.75	41
14	0.84	0.84	0.84	61
15	0.89	0.86	0.88	58
Accuracy		0.69		360
Macro Avg	0.64	0.62	0.62	360
Weighted Avg	0.69	0.69	0.69	360

Şekil 16: Düzeltilmiş GBM Performans Tablosu

Düzeltilmiş GBM modelin performans tablosu değerlendirildiğinde genel doğruluk accuracy değeri 0.69 çıktığı gözlemlenmiştir. Bu değerlere dayanarak modelin performansının ortalamanın olduğu söylenmektedir. Performansının 0.60 olması sebebi ile modelide düzeltme yapılmaştır. Yapılan düzeltmenin ilkel performans skorunu minimum derecede artırmıştır.

3.3.9 CatBoost

CatBoost sınıflandırma yöntemi kullanılarak bir model oluşturuldu.Sınıflandırma sonucunda modelin başarısını ve performansını değerlendirmek için sınıflandırma skoruna bakılmıştır.

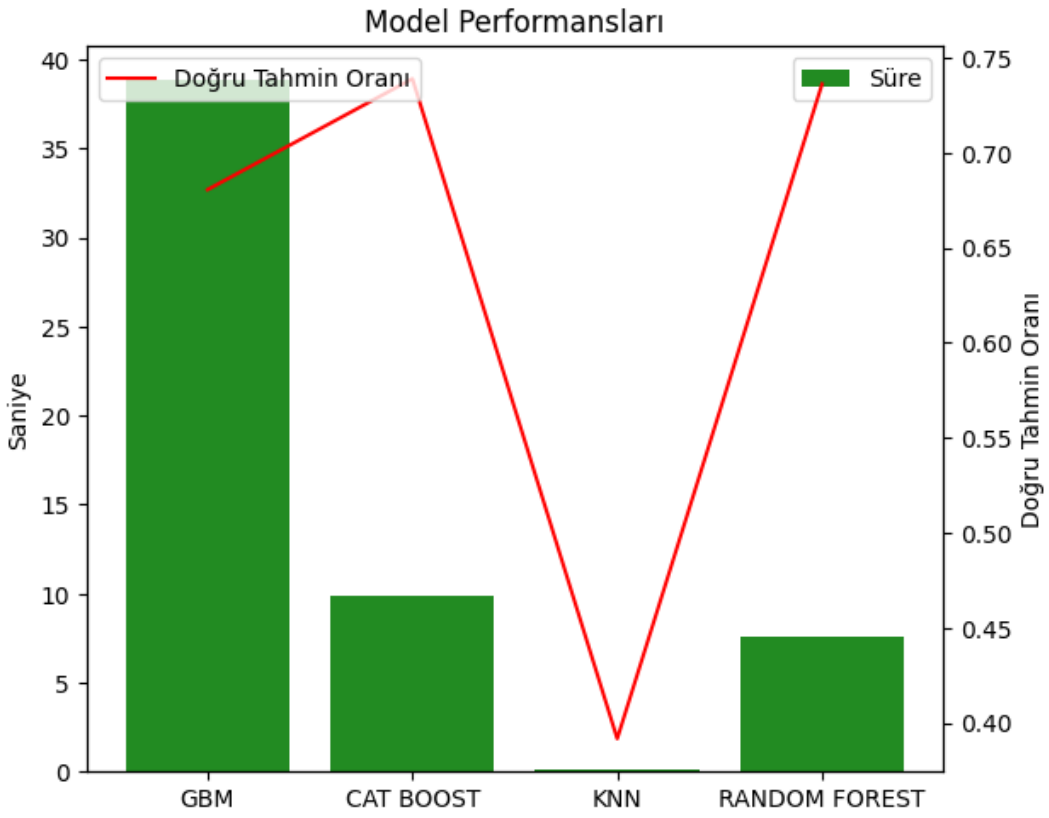
Class	Precision	Recall	F1-Score	Support
1	1.00	0.70	0.82	10
2	0.69	0.90	0.78	10
3	0.73	0.80	0.76	10
4	0.50	0.20	0.29	10
5	0.50	0.60	0.55	10
6	0.50	0.50	0.50	10
7	0.56	0.50	0.53	10
8	0.50	0.40	0.44	10
9	0.52	0.67	0.58	21
10	0.53	0.47	0.50	19
11	0.79	0.65	0.71	40
12	0.71	0.80	0.75	40
13	0.77	0.90	0.83	41
14	0.91	0.84	0.87	61
15	0.90	0.93	0.92	58
Accuracy		0.75		360
Macro Avg	0.67	0.66	0.66	360
Weighted Avg	0.75	0.75	0.74	360

Şekil 17: CatBoost Performans Tablosu

Performans tablosu değerlendirildiğinde genel doğruluk accuracy değeri 0.75 çıktığı gözlemlenmiştir. Bu değerlere dayanarak modelin performansının ortalama üstünde olduğu söylenmektedir. CatBoost yapısında modelin en performanslı olduğu metriklerin kendiliğinden optimize ettiği bilinmektedir. CatBoost'un bu özelliği nedeni ile model üzerinde herhangi bir düzeltme işlemi uygulanmamıştır.

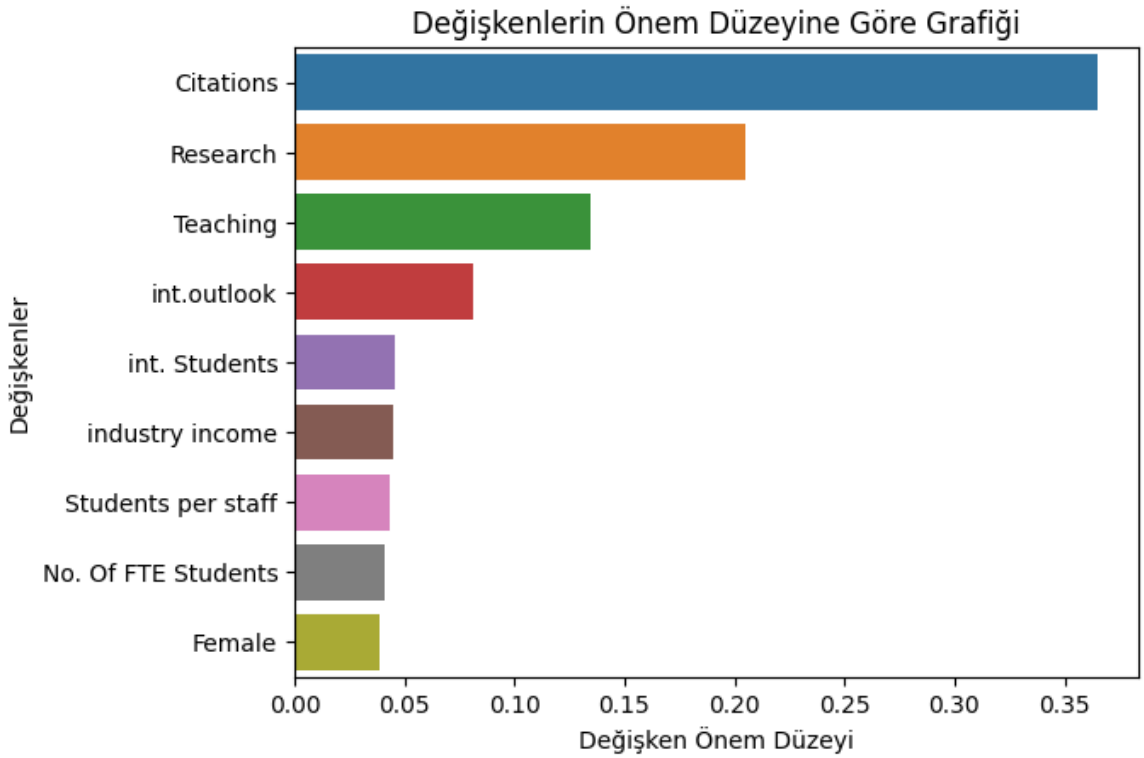
3.3.10 Performans Grafiği ve Değişken Önem Düzeyi

Veri setini en iyi şekilde analiz edebilmek için toplamda 10 adet model oluşturulmuştur. Modellerin içerisinde en iyi şekilde başarı gösteren Random Forest ve CatBoost modelleri olmuştur. Bu 2 modelin genel doğruluk (accuracy) skorularına bakıldığında araların sadece 0.01'lik bir değişim olduğunu gözlemlenmiştir. Senaryo testlerine geçmek için bir modelin belirlenmesinde performans değerlendirme grafiğine göz atılmıştır.



Şekil 18: Modellerin Performans Grafiği

Performans grafiğine göz atıldığında, dört farklı modelin süre ve doğruluk oranlarını gösterdiği görülmüştür. Model tercihi yapılırken, en kısa sürede işlem yapan ve yüksek doğruluk oranına sahip olan model tercih edilmiştir. Model seçimi yapıldıktan sonra değişken önem düzeyine bakılmıştır.



Şekil 19: Değişken Önem Düzeyi Grafiği

Değişken önem düzeyinde en önemli olan dört değişkenin Citations, Research, Teaching ve int.outlook olduğu görülmektedir. Senaryo adımlarında bu dört değişkenin özellikle değerlendirilmesi gerektiği anlaşılmaktadır.

3.4 Senaryo

Belirlenen model ile tahminleme işlemlerine başlamadan önce doğruluğundan emin olmak amacıyla Dokuz Eylül üniversitesinin ham veri setindeki verileri farklı bir veri olarak gösterilmiş ve modelden tahminlenmesi istenmiştir. Modelin tahminlediği sıra ile ham veri setindeki sırası ile uyuşmuştur.

```
deu_sıra=[14,56233,17.6,0.03,47,18.4,13.3,20.9,95.8,21.7]

uni=cat_final.predict(deu_sıra)

uni

array([14], dtype=int64)
```

Şekil 20: Model Tahmin Testi

Şekil 20'de görüldüğü gibi model Dokuz Eylül üniversitesinin sırasını 14. sıra grubu (1201-1500 arası) olarak bulmuş ve doğrulunu kanıtlamıştır.

CatBoost model ile senaryolara başlamak için 3 farklı veri setini hazırlanmış ve veri setindeki her bir değişken için rastgele değerler atanması koşulu ile toplamda 900 gözlem oluşturulmuştur.

İlk veri setinde her yüz gözlemde sadece bir değişken için rastlege değer atanmış ve modelin tahmin etmesi beklenmiştir. Modelin tahmin ettikten sonra tahminler incelenmiş ve Dokuz Eylül üniversitesinin maksimum dokuzuncu sıra grubuna çıktı görülmüştür. Üniversitenin dokuncu gruba çıkabilmesi için Citations(Atıf puanı) değerinin 100 puan olması, diğer değişkenlerin ham veri setindeki gibi

- **Student per staff:** 17.6
- **int. Student :** 0.03
- **Female:** 47
- **Teaching:** 18.4
- **Research:** 13.3
- **İndustry income:** 95.8
- **int. Outlook:** 21.7

olması gerektiği kararına varılmıştır.

Birinci veri setinin diğer senaryosu olan Dokuz Eylül Üniversitesinin 11'inci sıra grubuna (601-800 arası) yükselmek için yine Citations(atıf puanı) değişkeninde değişiklik olduğu görülmüştür. Üniversitenin 11'inci sıra grubuna girebilmesi için;

- **Student per staff:** 17.6
- **int. Student :** 0.03
- **Female:** 47
- **Teaching:** 18.4
- **Citations:** 67.86
- **Research:** 13.3
- **İndustry income:** 95.8
- **int. Outlook:** 21.7

puanlarına sahip olması gerektiği belirlenmiştir. Bu puanlar sağlandığı sürece Dokuz Eylül Üniversitesinin mevcut grubundan yükseleceği gözlemlenir.

İkinci veri setinde tüm değişkenlerin ve tüm gözlemlerin rastgele bir şekilde atanmaış ve modelin tahminlenmesi beklenmiştir. Model tahminleme işlemini tamamladıktan sonra Dokuz Eylül üniversitesinin dünya sıralamsında ki tahmini sırası brinci sıra grubunda olduğu görülmüştür. Üniversitenin birinci sıra grubuna (1-50 arası) girebilmesi için ,

- **Student per staff:** 21.03
- **int. Student :** 4.27
- **Female:** 63
- **Teaching:** 65.40
- **Research:** 64.27
- **Citations:** 86.69
- **İndustry income:** 18.27
- **int. Outlook:** 84.57

Sekiz değişkeninde asıl skorlarından daha yüksek skorlara sahip olması gerektiği gözlemlenmiştir.

Üçüncü veri setinde ise Değişken önem düzeyi grafiğinde gözlemlenmiş değişkenler üzerinde rastgele değerler verilmiş ve modelin tahminlemesi beklenmiştir. Modelin Tahminlediği sıralama kontrol edildiğinde Dokuz Eylül üniversitesinin birinci sıra grubuna (1-50 arası) çıktığı gözlemlenmiştir. Üniversitenin birinci sıra grubuna (1-50 arası) çıkabilmesi için Teaching, Research, Citations ve industry outcome değişkenlerinin puanları

- **Student per staff:** 17.6
- **int. Student :** 0.03
- **Female:** 47
- **Teaching:** 78.50
- **Citations:** 70.62
- **Research:** 73.50
- **İndustry income:** 95.8
- **int. Outlook:** 21.7

gibi olaması ve diğer değişkenlerin ham veri setinde ki ile aynı tutulması gerekmektedir.

4 Sonuç

Sonuç olarak Dokuz Eylül Üniversitesi'nin dünya sıralamasındaki durumu değerlendirilmiş ve üniversitenin sıralamasını daha iyi konumlara getirmek için makine öğrenmesi algoritmaları kullanarak modeller oluşturulmuştur. Elde edilen sonuçlar, üniversitenin mevcut sırasını daha iyi tahmin etmek ve stratejik planlar yapmak açısından önemlidir.

Belirlenen model, üniverisitenin dünya sıralamsını belirlemede kullanılan çeşitli faktörleri dikkate alarak sıralama tahminleri yapmıştır. Bu tahminler üniversitenin potansiyel sıralama değişimlerini anlamamıza yardımcı olmuştur.

Elde edilen bulgular ışığında, Dokuz Eylül Üniversitesi'nin dünya sıralamasındaki konumunu iyileştirmek için yapılacak stratejik planların önemli olduğu görülmektedir. Bu rapor, üniversitenin uluslararası alanda daha iyi bir konuma gelmesi, akademik başarılarını arttırması için değerlidir.

5 Kaynakça

Şahinaslan, Ö., Dalyan, H., Şahinaslan, E., (2022), Naive Bayes Sınıflandırıcısı Kullanarak YouTube Verileri Üzerinde Çok Dilli Duygu Analizi,*Bilişim Teknolojileri Dergisi, 15*,221-229

Chein, C. F., Chen, L. F. (2008) Data Mining to Improve Personnel Selection and Enhance Human Capital: A Case Study in High-Technology Industry, Expert Systems with Applications, vol. 34, p. 280-290.

Aytekin, Ç., Sütcü, C.S., Özfidan, U. (2018) Text Classıficatıon Vıa Decısıon Trees Algorıthm: Customer Comments Case, The Journal of International Social Research, vol:11 ss:55.

L. Breiman, "Random Forests", Machine Learning, 45(1), 5-32, 2001.

Y. Zhao, Y. Zhang, "Comparison of Decision Tree Methods for Finding Active Objects", Advances in Space Research, 41(12), 1955-1959, 2008.

Mitchell, T., "Machine Learning", McGraw Hill, New York, (1997).

Coomans, D and Massart, D.L., "Alternative k-nearest neighbour rules in supervised pattern recognitition: Part 1. Knearest neighbour classification by using alternative voting rules", Analytica Chimica Acta, 136: 15-27 (1982).

Aha, D.W., Kibler, D. and Goldstone, R.L., "Instance-based learning algorithms", Machine Learning, 6:37-66 (1991).

Ulupınar, S. D. (2007); 2001 Kriz Dönemi, Öncesi ve Sonrasında Türk Ticari Bankalarının Karlılıklarının Lojistik Regresyon Analizi ile İncelenmesi, İstatistik Bilim Dalı Yüksek Lisans Tezi, Marmara Üniversitesi, İstanbul.

Uğur, A., Kınacı, A. C. (2006). Yapay zeka teknikleri ve yapay sinir ağları kullanılarak web sayfalarının sınıflandırılması. XI. Türkiye'de İnternet Konferansı (inet-tr'06), Ankara, 1(4).

Alshari, H., Saleh, A. Y., Odabaş, A. (2021). Comparison of gradient boosting decision tree algorithms for CPU performance. Journal of Institue Of Science and Technology, 37(1), 157-168.