İLERİ İSTATİSTİK PROJELERİ(IST491)

May 27, 2021

HACETTEPE ÜNİVERSİTESİ

FEN FAKÜLTESİ

İSTATİSTİK BÖLÜMÜ

MAKİNE ÖĞRENİMİNDE SINIFLANDIRMA ALGORİTMALARININ KIYASLANMASI

Hazırlayan

Yasin KOÇAK

21722041

Dersin Sorumlusu

Doç. Dr. DERYA KARAGÖZ

ANKARA

2021

Teşekkürler

Bu projeye başlamam için bana bu fırsatı sunan, beni her konuda teşvik eden Doç. Dr. Derya KARAGÖZ hocama ve her sorumda yanımda olan, bana yol gösteren Arş. Gör. Dr. Mustafa Murat Arat hocama teşekkür ederim.

İçindekiler

1 Giriş

Yüksek lisans programlarına olan başvuruların kabulünü belirlemek için birçok kriter ele alınır.Bu projenin amacı öğrencinin herhangi bir yüksek lisans başvuru süreci için ortak giriş koşulları olarak kabul edilen kriterlere dayalı sınıflandırma yapmaktır. Veri seti 500 gözlem 9 değişkenden oluşmaktadır. Bu araştırma için kullandığım veri seti https://www.kaggle.com/mohansacharya/graduate-admissions adresinden alınmıştır.

DEĞİŞKENLER:

- 340 üzerinden değerlendirilen GRE puanı.
- 120 üzerinden değerlendirilen TOEFL puanı.
- ÜNİVERSİTE dereceleri. (1-5 arası puanlandırılmış, kategorik ve sıralı değişkendir).
- Amaç Beyanı (1-5 arası puanlandırılmıştır).
- Tavsiye Mektubu (1-5 arası puanlandırılmıştır).
- CGPA (akademik not ortalaması 6.8- 9.92 arası puanlandırılmıştır).
- Araştırma yapıp yapmama durumu.(1:araştırma yaptı, 0:araştırma yapmadı).
- Kabul şansı (0-1 arasında değerlendirilmiştir).

2 Veri Önişleme

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 500 entries, 0 to 499
Data columns (total 9 columns):
 #   Column             Non-Null Count  Dtype
---  ------             --------------  -----
 0   Serial No.         500 non-null    int64
 1   GRE Score          500 non-null    int64
 2   TOEFL Score        500 non-null    int64
 3   University Rating  500 non-null    int64
 4   SOP                500 non-null    float64
 5   LOR                500 non-null    float64
 6   CGPA               500 non-null    float64
 7   Research           500 non-null    int64
 8   Chance of Admit    500 non-null    float64
dtypes: float64(4), int64(5)
memory usage: 35.3 KB
```

Figure 1: Kayıp Gözlem

Sınıflandırma algoritmaları için seri numaraları işimize yaramayacağı için veri setinden çıkardım. Tablo 1'de incelendiğinde kayıp gözlem kontrolünde hiçbir kayıp gözleme rastlanılmamıştır.

	Serial No.	GRE Score	TOEFL Score	University Rating	SOP	LOR	CGPA	Research	Chance of Admit
count	500.000000	500.000000	500.000000	500.000000	500.000000	500.00000	500.000000	500.000000	500.00000
mean	250.500000	316.472000	107.192000	3.114000	3.374000	3.48400	8.576440	0.560000	0.72174
std	144.481833	11.295148	6.081868	1.143512	0.991004	0.92545	0.604813	0.496884	0.14114
min	1.000000	290.000000	92.000000	1.000000	1.000000	1.00000	6.800000	0.000000	0.34000
25%	125.750000	308.000000	103.000000	2.000000	2.500000	3.00000	8.127500	0.000000	0.63000
50%	250.500000	317.000000	107.000000	3.000000	3.500000	3.50000	8.560000	1.000000	0.72000
75%	375.250000	325.000000	112.000000	4.000000	4.000000	4.00000	9.040000	1.000000	0.82000
max	500.000000	340.000000	120.000000	5.000000	5.000000	5.00000	9.920000	1.000000	0.97000

Figure 2: Tanımlayıcı İstatistikler

Tablo 2'de bulunan tanImlayıcı istatistiklere bakıldığında yüksek lisans kabul şanslarının %75'i, %82 kabul şansından az olduğu görülmüştür. Hedef değişkenimiz olan yüksek lisans kabul şansını %82 den büyük ve küçük olacak şekilde (binary olarak) etiketledikten sonra sınıflandırma algoritmalarımızı kullanmaya devam edeceğiz.

2.1 Kutu Grafikleri İncelenmesi

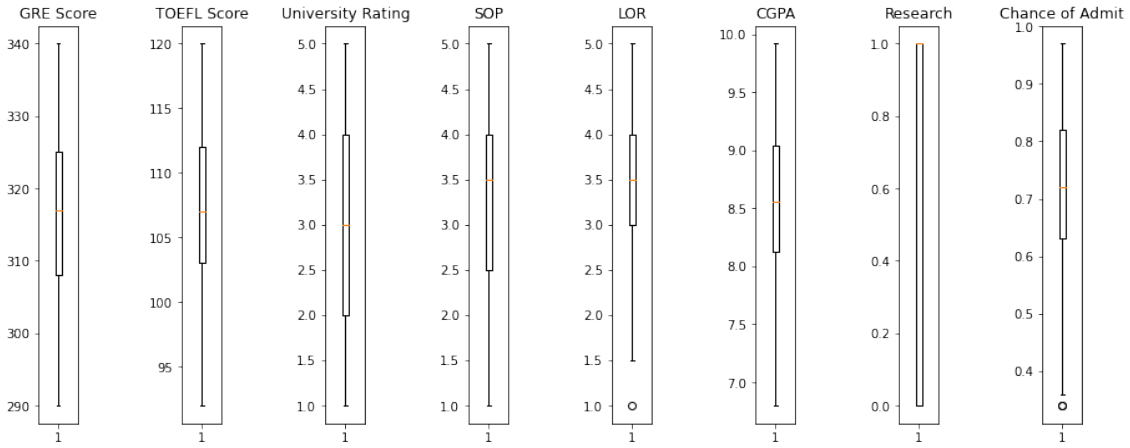

Figure 3: Kutu Grafikleri

Tablo 3'de bulunan kutu grafiklerinden, değişkenlerin ölçeklerinin benzer olmadığı ve sınıflandırma analizimizi çalıştırmadan önce normalleştirilmesi gerektiğini anlıyoruz. Ayrıca sınıflandırma analizimizi etkileyebilecek aykırı değerleride kontrol ettik. **Tavsiye mektubu ve kabul şansı** değişkenlerinin aykırı değerlere sahip olduğunu gözlemledik, ancak değerler beklenen ölçeğe uyduğundan veri giriş hataları gibi görünmediğinden ve bunları ortalama değerlerle ilişkilendirmek analizimizi önyargılı hale getireceğinden, aykırı değerler görmezden gelinmiştir.

2.2 Kategorik Değişkenlerin İncelenmesi

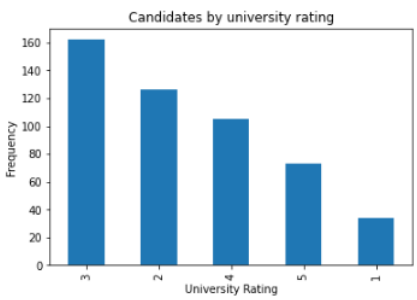

Figure 4: Üniversite Derecesi

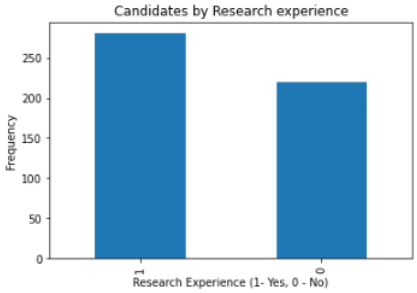

Figure 5: Araştırma Durumu

Yüksek lisans başvurusu en çok olan 3. ve 2. derece üniversitelere olmuştur. Yüksek lisans başvurusu yapan öğrenciler arasından araştırma deneyimi olan öğrenci sayısı daha fazladır.

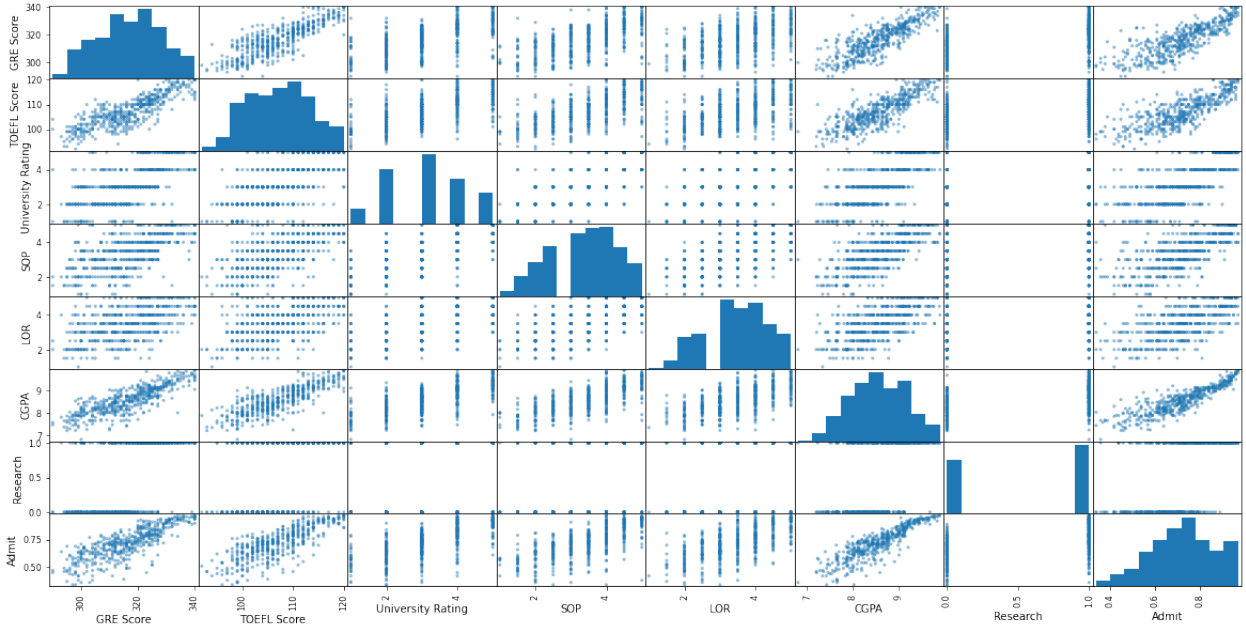

Figure 6: Dağılım Grafikleri

Tablo 6'ya bakıldığında GRE puanı TOEFL puanı ve CGPA akademik not ortalaması kabul şansıyla pozitif bir doğrusal ilişki gözlenir. Üniversite derecesi yükseldikçe kabul şansının da arttığı dağılım grafiğinde görülür. Araştırma deneyimine bakıldığında belirgin bir ilişki görülmesede araştırma deneyimi olan öğrencilerin yükek lisansa kabul şansı biraz daha yüksek olduğu görünüyor.

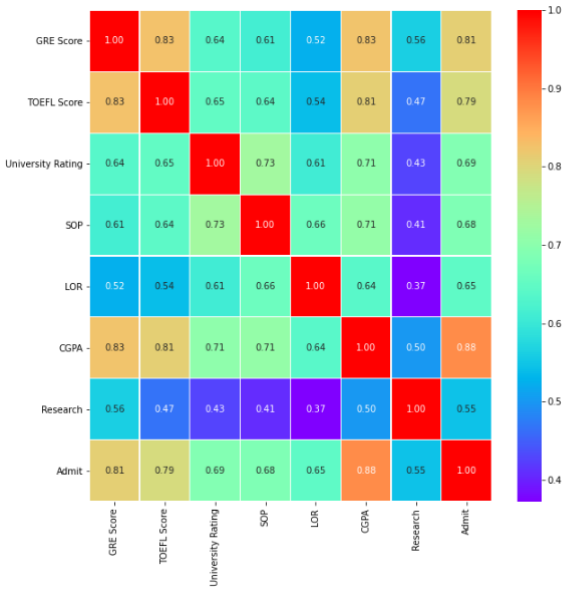

Figure 7: Korelasyon katsayıları

Tablo 7'ye bakıldığında CGPA akademik not ortalaması %88 GRE puanı %81 TOEFL puanı %79 luk bir ilişki kat sayısı görülmüştür.

2.3 Çok Değişkenli Görselleştirme

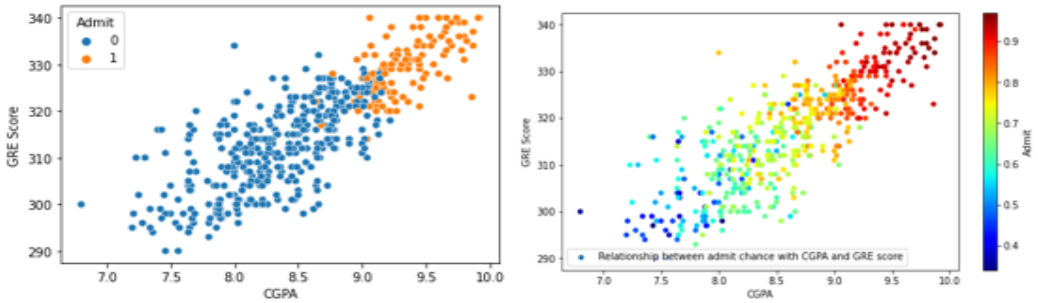

Figure 8: Toefl-Gre Puanları

Tablo 8'de belirtilen grafiklerde de beklendiği gibi, yüksek GRE puanlarına ve CGPA'ya sahip adayların kabul şansını doğrudan etkilediği daha da doğrulanabilir. Birinci grafikte sınıflandırma algoritmalarında kullanacağımız şekilde kabul şanslarının %82 den büyük olanların kabul olacağı düşük olanların reddedileceği şekilde etiketlenerek grafik çizdirilmiştir. İkinci grafikte de görüleceği gibi GRE puanında ve CGPA akademik not ortalamasındaki artış yüksek lisansa kabul şansında artışa neden olmuştur.

3 Veriyi Modelleme

Bu sınıflandırmanın bir parçası olarak, kabul şansı %82'den yüksekse öğrencinin yüksek lisans başvurusunun kabulü, düşükse başvurusunun reddedilmesi şeklinde hedef değişkenini oluşturup veri setimizden hedef değişken olarak kabul şansı değişkenini çekiyoruz. Kategorik değişken olan Üniversite Dereceleri dummy değişken olarak "one hot encoding" yöntemiyle etiketlenmiştir. Daha sonrasında sayısal gözlemlerimizde farklı ölçekler kullanıldığı için 0 la 1 arasında normalleştirme uygulanmıştır.

Veri setimiz sınıflandırma için hazırdır. Kullanmak için sahip olduğumuz veri kümesi 500 adet gözlem içerir, bu nedenle tüm veri kümemizi hiperparametre ayarı ve performans karşılaştırmaları için kullanırız. Mevcut tüm verileri tabakalandırma kullanarak %70 %30 oranıyla eğitim ve test seti halinde böldük.

3.1 Özellik Seçimi 'Boyutsal Küçültme'

Özellik seçimi, bir hedef değişkeni tahmin etmede ne kadar yararlı olduklarına bağlı olarak girdi özelliklerine bir puan atayan teknikleri ifade eder. Özellik önem puanlarının birçok türü ve kaynağı vardır, ancak popüler örnekler arasında istatistiksel korelasyon puanları, doğrusal modellerin parçası olarak hesaplanan katsayılar, karar ağaçları ve permütasyon önem puanları yer alır.

Rasgele ormanlar sınıflandırma algoritmasını kullanarak özellik seçimi kullanılmıştır.

- GRE Score = 0.166008088725173
- TOEFL Score = 0.17567262567024367
- SOP = 0.11266617434635785
- LOR = 0.06674586063885228
- CGPA = 0.36853842135277765

LOR Tavsiye Mektubu değişkeninin modeli açıklama puanı düşük olduğundan sınıflandırma algoritmalarını kullanırken LOR değişkeni çıkartılmış veri setiyle devam edilecektir.

:

	GRE Score	TOEFL Score	SOP	CGPA	Research	University Rating_1	University Rating_2	University Rating_3	University Rating_4	University Rating_5
0	337	118	4.5	9.65	1	0	0	0	1	0
1	324	107	4.0	8.87	1	0	0	0	1	0
2	316	104	3.0	8.00	1	0	0	1	0	0
3	322	110	3.5	8.67	1	0	0	1	0	0
4	314	103	2.0	8.21	0	0	1	0	0	0

4 Performans Ölçütleri

Etiketlenmiş veri, etiketlenmemiş verilerin her bir parçasını bir şekilde bilgilendirici veya bilinmesi istenen anlamlı özelliklerle etiketlenmiş halidir.Denetimli öğrenme ise hakkında öngörüde bulunmak istediğim bir değişken için yapay zeka algoritması kullanıyorsam ve bu algoritma tahmin edeceğim değişken için daha önceki gözlemleri kullanıyorsa denetimli öğrenme gerçekleştiriliyordur. Etiketlenmiş veriler kullanılır. Makine öğrenimi süreçlerinden biri olan sınıflandırma algoritmaları denetimli öğrenmedir. Sınıflandırma, fikirleri ve nesneleri önceden belirlenmiş kategorilere veya 'alt popülasyonlara' tanıma, anlama ve gruplama sürecidir. Önceden kategorize edilmiş eğitim veri kümelerini kullanan makine öğrenimi programları, gelecekteki veri kümelerini kategorilere ayırmak için çeşitli algoritmalar kullanır.

Bu bölümde genel olarak sınıflandırıcıları değerlendirmek için kullanılan ölçütlerden bahsedilecektir.

4.1 Karışıklık Matrisi

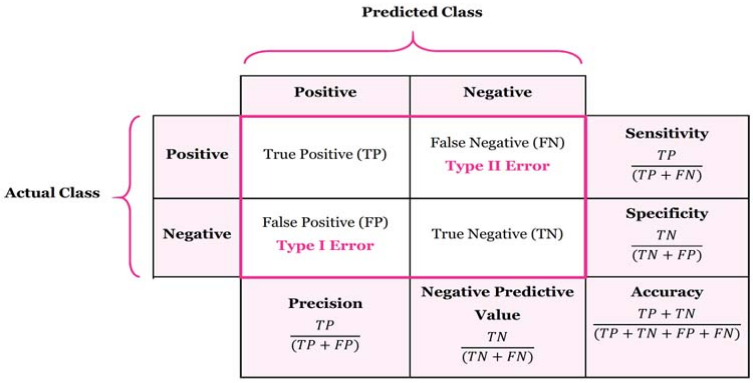

Figure 9: Karışıklık Matrisi

VERİ setimize göre;

- **TP:** yüksek lisans başvuruları kabul edilen öğrencileri kabul edilmiş şekilde tahmin etmek.
- TN: yüksek lisans başvuruları reddedilen öğrencilerin reddedildiğini tahmin etmek.
- FP: yüksek lisans başvuruları reddedilmiş öğrencilerin kabul edildi olarak tahmin etmek.
- FN: yüksek lisans başvuruları kabul edilmiş öğrencilerin reddedildi olarak tahmin etmek.
- Accuracy(tutarlılık): Doğru tahminlerin tüm tahminlere oranıdır.(eşit dağılmayan veri setlerinde pek kullanışlı değildir).
- Sensitivity-recall(duyarlılık):Doğru tahmin edilen pozitiflerin, doğru tahmin edilen pozitiflerin ve yanlış tahmin edilen negatiflere oranıdır.
- Specificity(özgüllük): Doğru tahmin edilen negatiflerin tüm negatifl tahiminlere oranıdır.
- Precision(kesinlik): Doğru tahmin edilen pozitiflerin, tüm pozitiflere oranıdır.
- F1 puanı: Kesinlik ve duyarlılık arasındaki dengeyi açıklar.

4.2 Roc Eğrisi ve Eğri Altında Kalan Alan(ROC(Receiver Operating Characteristic) CURVE-AUC)

Accuracy (tutarlılık) den sonra en çok kullanılan performans ölçüsüdür. TPR(Doğru pozitif oranı): duyarlılık diyebiliriz, doğru pozitif tahminlerin tüm pozitif tahminlere oranıdır. FPR(Yanlış pozitif oranı): yanlış pozitif tahminlerin tüm negatif tahminlere oranıdır.

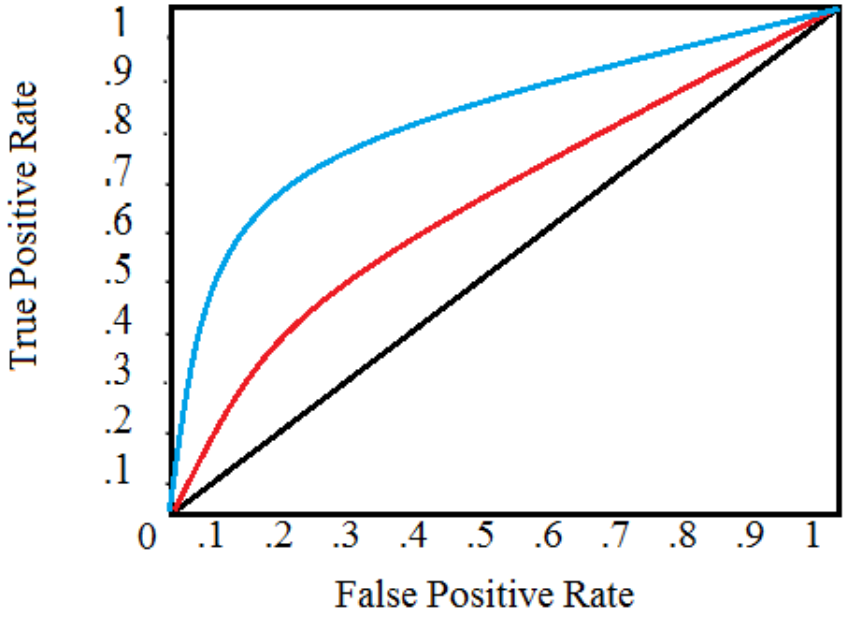

Figure 10: Auc

Sınıflandırma performansımı öğrenmek için roc çizgisinin altında kalan alana göre ölçülür, bu ölçü birimine de auc puanı denir.Auc puanının yüksek olması sınıflandırma algoritmamımızın ne kadar iyi çalıştığını gösterir (overfitting dışında). TPF nin yüksek FPR nin düşük olması beklenir.

Overfitting (aşırı uyum):Eğitim verilerini çok iyi modelleyen eğitim verisini ezberleyen modeli ifade eder. Bu durumda eğitim seti dışında karşılaşacağı bir modellemede aynı performansı gösteremeyecektir, bu durumda overfitting ortaya çıkacaktır.

Underfitting(yetersiz uyum):Eğitim verilerini modelleyemeyen veya yeni verilere genelleştiremeyen bir modeli ifade eder.

5 Sınıflandırma Algoritmaları Classification Algorithms

5.1 K En Yakın Komşu Algoritması (KNN, K-Nearest Neighbor Algorithms)

K sayıda komşu sayısı belirleyip, komşular arası mesafeyi hesaplayıp, en yakın komşuları bulup, komşularını etiketleyen bir sınıflandırma algoritmasıdır.

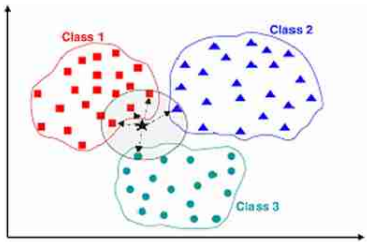

Figure 11: Knn

KNN algoritması için kullandığım hiperparametrelerden bahsetmeden önce grid search (ızgara araması) ve cross validation (çapraz doğrulama) dan bahsedelim.

Izgara Araması (Grid Search): Hiperparametre değerlerinin birlikte en yüksek auc puanını veren kombinasyonunu bulmak için tüm hiperparametre kombinasyonlarını arayan bir yöntemdir. Izgara araması için gereken, kullanacağımız parametreler ve hangi değerleri alacağını belirlemektir.

Çapraz Doğrulama (Cross Validation): Çapraz doğrulama, sınırlı bir veri örneğinde makine öğrenimi modellerini değerlendirmek için kullanılan bir yeniden örnekleme prosedürüdür.

Prosedürün, belirli bir veri örneğinin bölüneceği grupların sayısını ifade eden k adında tek bir parametresi vardır. Bu nedenle, prosedür genellikle k-kat çapraz doğrulama olarak adlandırılır.

Figure 12: Çapraz Doğrulama

KNN Hiperparametreleri;

1. **n_neighbors (komşu sayısı):**
   - **weights (ağırlıklar): tahminde kullanılan ağırlık fonksiyonu. Olası değerler:**
   - **"uniform": tek tip ağırlıklar. Her mahalledeki tüm noktalar eşit olarak ağırlıklandırılmıştır.**
   - **"distance": mesafelerinin tersine göre ağırlık noktaları. Bu durumda, bir sorgu noktasının daha yakın komşuları, uzaktaki komşulardan daha büyük bir etkiye sahip olacaktır.**
2. **P(mesafeler):**
   - **Euclidean(Öklid mesafesi): Matemetikte pisagor bağlantısı kullanılarak bulunan iki nokta arasındaki mesafe ölçüm birimidir.**
     $d(i,j) = \sqrt{\sum_{k=1}^{p}(x_{ij} - x_{jk})^2}, i,j = 1,2...n; k = 1,2...p$
   - **Manhattan: Hesaplama yöntemi olarak verilen iki noktanın koordinatlarının farkının mutlak değeri kullanılabilir.**
     $d(i,j) = \sum_{k=1}^{p}(\mid (x_{ij} - x_{jk} \mid)), i,j = 1,2...n; k = 1,2...p$
   - **Minkowski:En çok kullanılan uzaklık fonksiyonu Minkowski uzaklık fonksiyonudur. Genel olarak Öklid uzaklığı, Manhattan city-blok uzaklığı ve ölçekli Öklid uzaklığı Minkowski fonksiyonundaki lamda'nın farklı değerlerinde elde edilen ölçülerdir.**
     $d(i,j) = (\sum_{k=1}^{p}(\mid (x_{ij} - x_{jk} \mid)^p)^{1/p}, i,j = 1,2...n; k = 1,2...p$

5.2 Karar Ağaçları Algoritması (Decision Tree Algorithms)

Karar ağaçları hiyerarşik bir yapıya sahip, başta kök düğümle başlayıp yaprak düğümlere ayrılan ve bu ayrımları en yüksek bilgi kazanımına göre yapan aç gözlü bir algoritmadır.

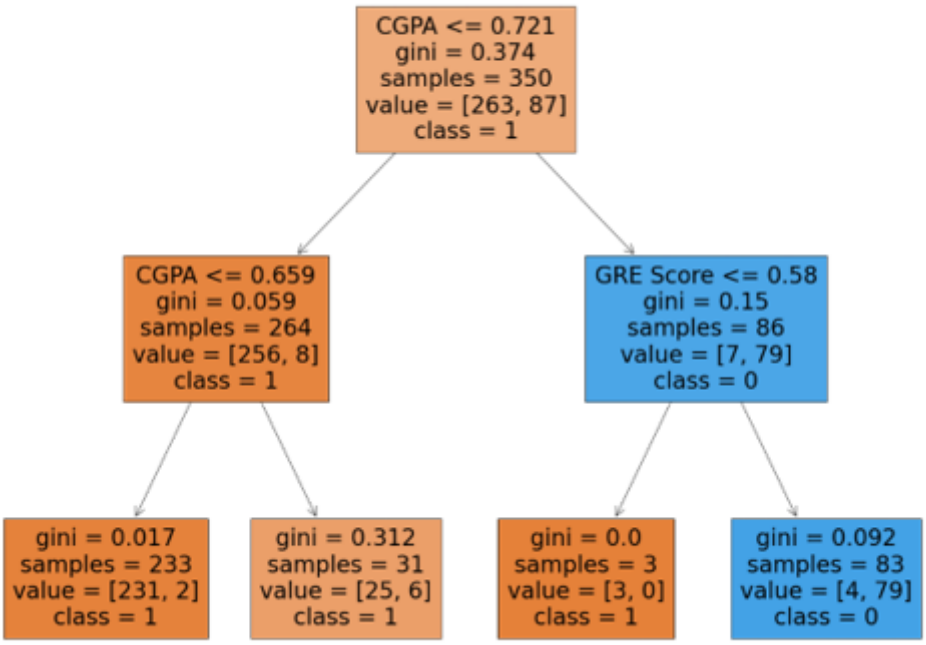

Figure 13: Karar Ağacı

Bilgi Kazanımı (Information Gain) : Bir karar algoritması kullanılarak bilgi kazanımı en üst seviyeye çıkaran özellik üzerinde bölerek standart sapmaların azaltılması sağlanılıyor. entropy ve gini den bahsedelim.

$entropy = -plog_2(p) - qlog_2(q)$

$gini = 1 - \sum p_j^2$

Karar ağaçlarında bir düğümü bölmek için kullanılan bilgi kazanımı hesaplama metotlarıdır. Genellikle benzer karar ağacı oluştururlar.

Karar Ağaçları Hiperparametreleri

1. **Criterion: entropy gini**
2. **Splitter: Her düğümde bölünmeyi seçmek için kullanılan stratejilerdir. En iyi bölmeyi seçmek için, best "en iyi" ve en iyi rasgele bölmeyi seçmek için,random "rasgele" stratijileridir.**
3. **max depth :Ağacın maksimum derinliği. Yok ise, düğümler tüm yapraklar saf olana genişletilir.**

5.3 Rastgele Ormanlar Algoritması (Random Forest Algorithms)

Adındanda anlaşılacağı gibi bir topluluk olarak çok sayıda karar ağacından oluşur.Rastgele ormandaki her bir ağaç bir sınıf tahmini verir ve en çok oyu alan alan sınıf modelimizin tahmini haline gelir.

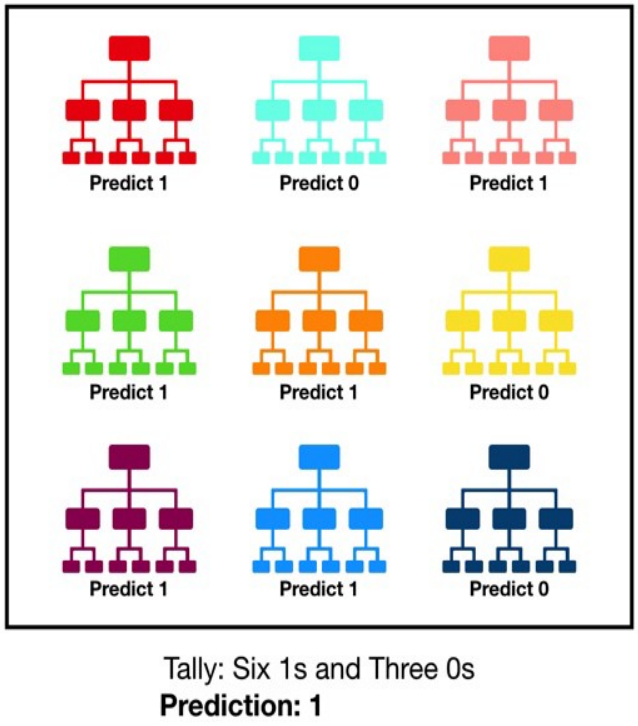

Figure 14: Rastgele Ormanlar

Random Forest Hiperparametreleri Karar ağaçları ile aynı hiperparametreler kullanılmıştır. Sadece random forest a özel olan 'n estimators' parametresi kullanılmıştır.Bu parametre algoritmada kaç tane karar ağacı kullanılacağından bahseder.

5.4 Lojistik Regresyon(Logistic Regression)

Lojistik regresyon, bir sonucu belirleyen bir veya daha fazla bağımsız değişken bulunan bir veri kümesini analiz etmek için kullanılan istatistiksel bir yöntemdir. Sonuç, ikili bir değişkenle ölçülür (yalnızca iki olası sonuç vardır). Doğrusal sınıflandırma problemlerinde yaygın bir biçimde kullanılır Lojistik regresyonun amacı, iki yönlü karakteristiği (bağımlı değişken = yanıt veya sonuç değişkeni) ile ilgili bir dizi bağımsız (öngörücü veya açıklayıcı) değişken arasındaki ilişkiyi tanımlamak için en uygun (henüz biyolojik olarak makul) modeli bulmaktır. Lojistik regresyon varsayılan sınıfın olasılığını modeller. Bağımlı değişken kategorik ikili ya da daha fazla değer aldığından basit doğrusal regresyondan ayrılır.

5.5 Destek Vektör Makineleri (Support Vector Machine Algorithms

İki veri noktası sınıfı ayırmak için seçilebilecek birçok olası hiper düzlem vardır. Amaç her iki sınıfın beri noktaları arasındaki maksimum mesafeyi bulmaktır. Maksimum mesafe gelecekteki veri noktalarının daha güvenli bir şekilde sınıflandırılabilmesi için güçlendirme sağlar.

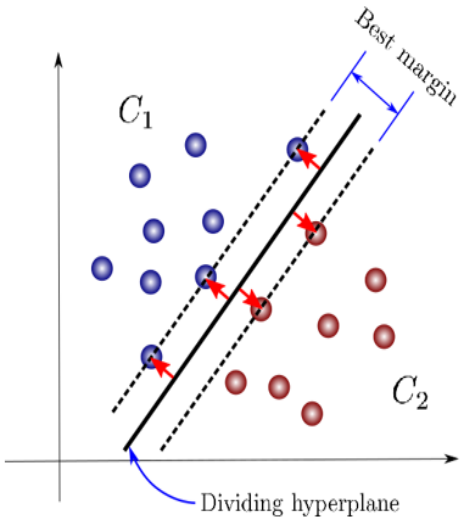

Figure 15: SVM

SVM Hiperparametreleri

1. **kernel: Algoritmada kullanılacak çekirdek türünü belirtir. "Doğrusal", "poli", "rbf", "sigmoid", "önceden hesaplanmış" veya çağrılabilir özelliklerden biri olmalıdır. Hiçbiri belirtilmezse, "rbf" kullanılacaktır.**

2. **C: Düzenlilik parametresi. Düzenlemenin gücü C ile ters orantılıdır. Kesinlikle pozitif olmalıdır. Ceza, kare şeklinde bir l2 cezasıdır.**

3. **Gamma: 'Rbf', 'poli' ve 'sigmoid' için çekirdek katsayısıdır.**

5.6 XG Boost

Toplu Öğrenme(Enseble Learning): Bir grup sınıflandırıcı algoritmanın bireysel olarak gerçekleştirdikleri performanslarından daha iyi bir performans göstermek için topluluk halinde çalıştıkları bir tür makine öğrenimi tekniğidir. Aşırı gradyan arttırma(XGBoost, eXtreme Gradient Boosting) anlamına gelir. Bu zayıf algoritmaları birleştirme yöntemlerinden xgboost, Boosting (arttıma) yöntemini kullanır Boosting; Genellikle homojen zayıf öğrenenler olarak kabul edilen destek, onları çok uyarlanabilir bir şekilde sırayla öğrenir (temel model öncekilere bağlıdır) ve bunları belirleyici bir strateji izleyerek birleştirir.
XGB Hiperparametreleri

1. **Min child weight: Aşırı öğrenmeyi (over fitting) kontrol etmek için kullanılır. Daha yüksek değerler, bir modelin, bir ağaç için seçilen belirli bir örneğe oldukça özel olabilecek öğrenme ilişkilerini engeller.**

2. **Eta: Her adımda ağırlıkları küçülterek modeli daha sağlam hale getirir Kullanılacak tipik nihai değerler: 0,01-0,2**

3. **Max depth: Bir ağacın maksimum derinliği, GBM ile aynı. Daha yüksek derinlik, modelin belirli bir numuneye çok özel ilişkileri öğrenmesine izin vereceğinden aşırı uydurmayı kontrol etmek için kullanılır. CV kullanılarak ayarlanmalıdır. Tipik değerler: 3-10**

6 Sonuçların Karşılaştırılması

Bu bölümde kullanılan tüm algoritmaların bahsedilen performans ölçütlerinden AUC punanı ve Accuracy puanına göre karşılaştırmaları yapılmıştır.

Siniflandırma Algoritması;	KNN	DECISION TREE	RANDOM FOREST	LOGISTIC REGRESSION	SUPPORT VECTOR MACHINE	XGB
AUC;	0.833652	0.896795	0.856254484	0.8248026	0.8427409	0.8518297
accuracy	0.9	0.9266	0.9066	0.8866	0.9	0.9

- **KNN Algoritması Test Veri Seti İçin;Izgara araması yapıldığında en iyi parametreler 'nneighbors': 3, 'p': 1, 'weights': 'uniform' olarak bulunmuştur. En iyi parametreler acuracy puanı en yüksek olduğu zaman belirlenir. Test seti tahminlerine göre accuracy 0.9 puan, AUC 0.83 puan almıştır.**
- **Decision Tree Algoritması Test Veri Seti İçin;Izgara taraması yapıldıktan sonra en iyi parametreler 'criterion': 'gini', 'maxdepth': 2, 'splitter': 'best' belirlenmiştir. Test Seti tahminlerine göre auc puanı 0.89 accuracy puanı 0.926 puanı almıştır.**
- **Random Forest Algoritması Test Veri Seti İçin;Izgara taraması yapıldıktan sonra en iyi parametreler criterion: 'gini', max depth: 8, nestimators: 100 olarak belirlenmiştir. Test Seti tahminlerine göre auc puanı 0.85 accuracy puanı 0.9 puanı almıştır.**
- **Logistic Regression Algoritması Test Veri Seti İçin;Test Seti tahminlerine göre auc puanı 0.82 accuracy puanı 0.8 puanı almıştır.**
- **SVM Algoritması Test Veri Seti İçin;Izgara taraması yapıldıktan sonra en iyi parametreler 'C': 0.3, 'degree': 4, 'gamma': 'scale', 'kernel': 'poly' olarak belirlenmiştir. Test Seti tahminlerine göre auc puanı 0.84 accuracy puanı 0.9 puanı almıştır.**
- **XGB Algoritması Test Veri Seti İçin;Izgara taraması yapıldıktan sonra en iyi parametreler 'eta': 0.1, 'maxdepth': 3, 'min child weight': 1 olarak belirlenmiştir. Test Seti tahminlerine göre auc puanı 0.85 accuracy puanı 0.9 puanı almıştır.**

Tabloda görüldüğü üzere karar ağaçları en yüksek tutarlılık ölçüt sonucunu vermiştir.Gerçek hayattada yüksek lisans başvuru sonuçları belirli bir karar algoritması sonucu açıklanır.Karar ağaçları algoritmalarının çalışma mantığı incelendiğinde gerçek hayattaki gibi belirli bir mantık üzerine çalıştığı için bu algoritmanın en yüksek tutarlılığı vermiş olması beklenildiği gibi gerçekleşmiştir.

7 Referans

- **13.03.2017 Aurelien Geron Hands-On Machine Learning with Scikit-Learn, Keras**
- **figure10 https://manisha-sirsat.blogspot.com/2019/04/confusion-matrix.html**
- **figure11 https://www.datasciencearth.com/roc-egrisi-nedir/**
- **figure12 https://laptrinhx.com/chapter-1-k-nearest-neighbours-classifier-2516295810/**
- **figure13 https://medium.com/the-owl/k-fold-cross-validation-in-keras-3ec4a3a00538**
- **figure18 https://towardsdatascience.com/understanding-random-forest-58381e0602d2**
- **figure19 https://towardsdatascience.com/support-vector-machines-for-classification-fc7c1565e3**