Bil 470-Proje Raporu

Yüksek Öğrenim Öğrencileri Performans Değerlendirmesi

Zeynep Öztekin 191101044

1. Bilgisayar Mühendisliği Bölümü
Tobb Ekonomi ve Teknoloji Üniversitesi
zoztekin@etu.edu.tr

Proje No:63

1. Özet

2019 yılında "The Faculty of Engineering and Faculty of Educational Science" adlı okulun 145 farklı öğrencisinden toplanan 30 farklı girdi ile oluşturulmuş dataseti kullanılarak, Decision Tree ve Random Forest yapısı kullanılarak bu proje oluşturulmuştur.

2. Giriş

2.1. Motivasyon

Genel olarak tüm öğrenciler not durumu hakkında büyük bir meraka sahiptir. Bu projeden bazı sahip olunan girdiler ile öğrencinin aldığı dersteki başarı durumunu bulmayı amaçladım.

2.2. Amaç/ Hedef

Elimde bulunan dataset ile bir öğrencinin girilen verilere göre belirli bir derste hangi notu alacağını bulmayı hedefledim. Machine learning tekniklerinden olan Decision Tree ve Random Forest yapısını kullanarak makinenin yaklaşık doğru bir tahmin vermesini amaçladım. Bu projenin amacı öğrencinin performansını etkileyen faktörler ile öğrencinin dönem sonucunu tahmin edebilen makine öğrenmesi tabanlı bir sistem geliştirmektir.

3. Literatür araştırması

Seçtiğim veri setinin konusu çok yaygın olduğundan ötürü internette birçok kaynağa denk geldim. Kullandığım kaynakları referans bölümünde belirttim.

4. Veri Seti, Veri özellikleri, Öznitelikler

4.1.

https://www.kaggle.com/datasets/csafrit2/higher-education-students-performance-evaluation

Veri setinde öğrencinin yaşı, cinsiyeti, mezun olduğu lise tipi, ek iş yapma durumu, partnere sahip olma durumu, eğer varsa toplam geliri, üniversiteye ulaşımı, konaklama türü, anne ve babanın eğitim durumu, eğer varsa kardeş sayısı, anne babanın iş durumu gibi kişi hakkında genel bilgiler ve haftalık çalışma saati, okuma sıklığı ve okunan kitap türü, seminerlere katılma durumu, vizelere nasıl hazırlandığı, derslerde not alma durumu, dersi dinleme durumu, geçmiş dönem not durumu gibi kişinin direk eğitim ile alakalı girdileri bulunmaktadır.

Kullandığım veri setinde herhangi bir eksik veya hatalı veri durumu bulunmamaktadır. Projeyi gerçekleştirirken ham veri setini kullanabildim.

```
studentid       0
age             0
gender          0
hs_type         0
scholarship     0
work            0
activity        0
partner         0
salary          0
transport       0
living          0
mother_edu      0
father_edu      0
#_siblings      0
kids            0
mother_job      0
father_job      0
study_hrs       0
read_freq       0
read_freq_sci   0
attend_dept     0
impact          0
attend          0
prep_study      0
prep_exam       0
notes           0
listens         0
likes_discuss   0
classroom       0
cuml_gpa        0
exp_gpa         0
course id       0
grade           0
dtype: int64
```

	count	mean	std	min	25%	50%	75%	max
#_siblings	145.000000	2.806897	1.360640	1.000000	2.000000	3.000000	4.000000	5.000000
course id	145.000000	4.131034	3.260145	1.000000	1.000000	3.000000	7.000000	9.000000
grade	145.000000	3.227586	2.197678	0.000000	1.000000	3.000000	5.000000	7.000000

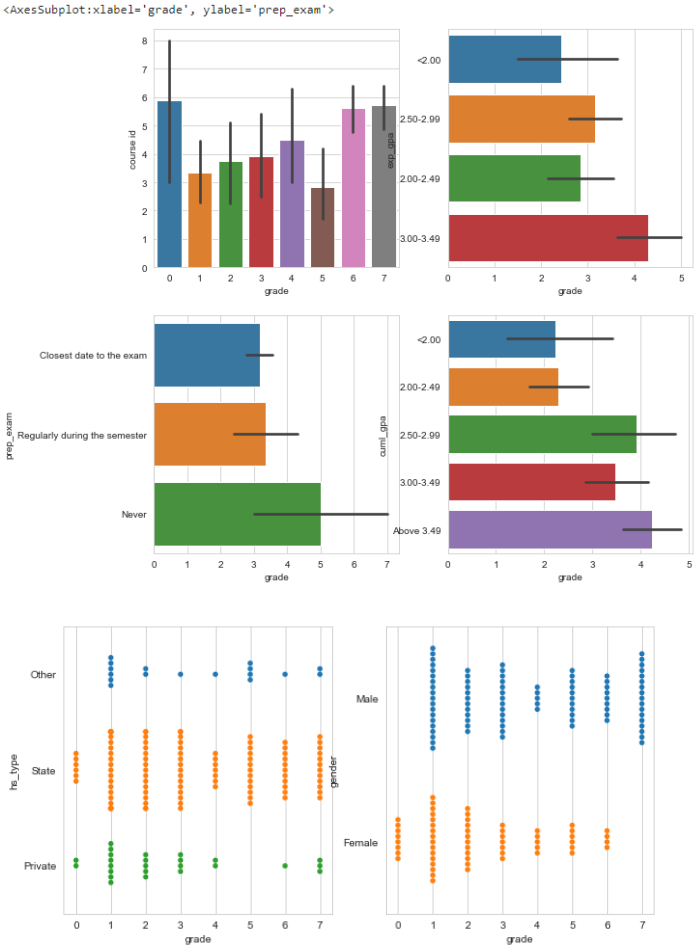
all feature is categorical, no missing value

Kullandığım veri setindeki bütün veriler sıralı ve karşılaştırılabilirdi.
Verilerim sayısal bir şekilde hazır haldeydi. Cinsiyet gibi sayısal tipte olmayan veriler sayısal hale dönüştürülmüştür (Erkek:1, Kadın:2 gibi)

Sonuç olarak alınan not durumu sıralı halde bulunmaktadır.

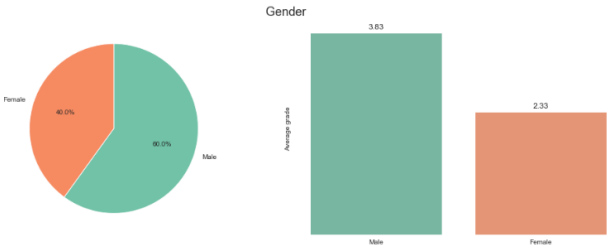
4.2.

Veriler önceden işlenmiştir ve kayıp değerler kontrol edilmiştir. Önerilen algoritmaların tahminde doğruluk açısından performansı, eğitimin çalışma süresi, tahminde bulunma ve her bir özelliğin sonuçlar üzerindeki etkisi gibi çeşitli faktörler göze alınmıştır.

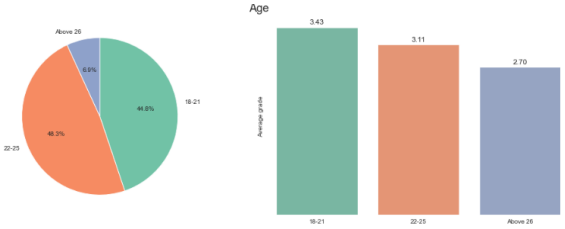


Student Id adlı kolonun sonuç üzerinde bir etkisi olmaması durumu sebebiyle kolonu çıkarttım. Yaptığım bazı işlemler sonucunda gözlemlediklerim;

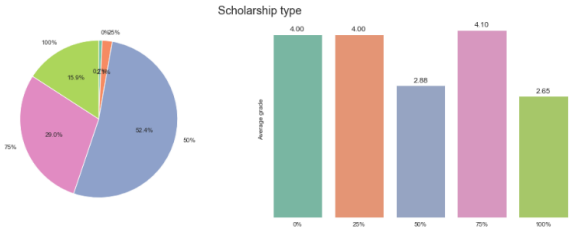
- Faculty of Engineering and Faculty of Educational Sciences okulunda kadın öğrencilerden çok erkek öğrenciler bulunmaktadır



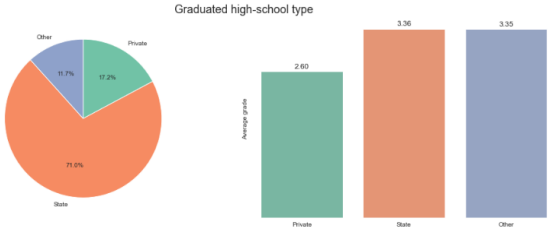
- Genç öğrencilerin daha yüksek notları olmaktadır



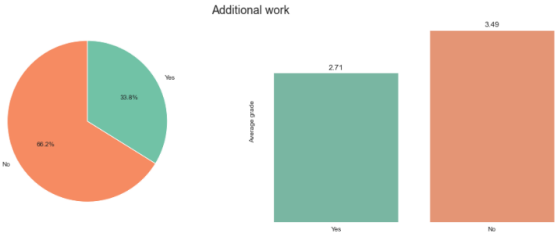
- Sahip olunan burs derecesinin öğrencinin başarı durumunda çok etkisi bulunmamaktadır



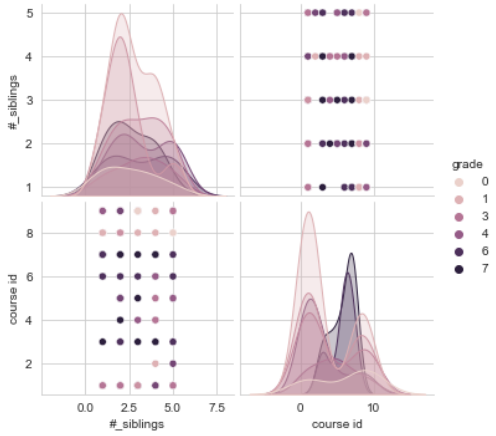
- Özel lise mezunu öğrencilerin diğer lise tiplerinde bulunan öğrencilere göre başarısı daha düşüktür



- Ek iş sahibi olmayan öğrencilerin ders başarı durumu daha yüksektir

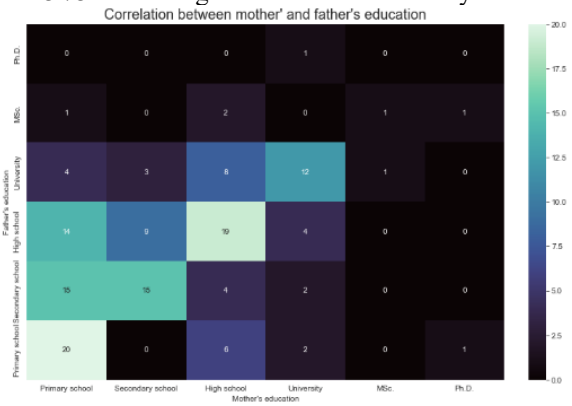


- Sahip olunan kardeş sayısı ve course id'nin sonuç ile korelasyonu

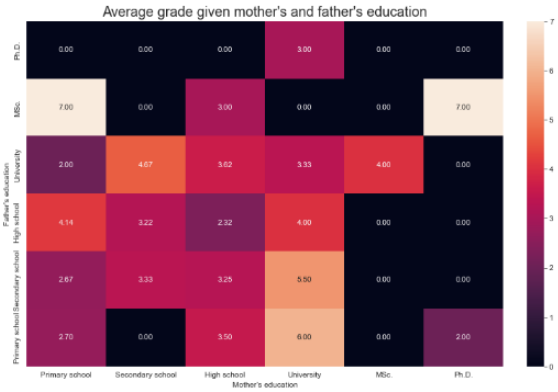


- Hiçbir öğrencinin beklenen GPA'sı 3.49 üstünde değil. Genel not ortalaması 2.5 ve üzeri olan öğrenciler, daha düşük GPA'a sahip olması beklenirken, daha düşük genel not ortalamasına sahip öğrencilerin daha yüksek bantta daha fazla beklenen GPA'a sahiptir.

- Anne ve babanın eğitimi arasındaki korelasyon



- Ebeveynlerin eğitim seviyesindeki fark, öğrencilerin notu üzerinde net bir etkisi bulunmamaktadır.



5. Kullanılan Modeller

Bu modeller, öğrencilerin performansını etkileyen faktörlerin bulunmasına ve performanslarının tahmin edilmesine yardımcı olur. Bu çalışmada, öğrenci performansı tahmin etmek için makine öğrenmesi algoritmaları kullanılır.

- **Decision Tree**

Bir karar ağacı, çok sayıda kayıt içeren bir veri kümesini, bir dizi karar kuralları uygulayarak daha küçük kümelere bölmek için kullanılan bir yapıdır. Basit karar verme adımları uygulanarak, büyük miktardaki kayıtları, çok küçük kayıt guruplarına bölerek kullanılan bir yapıdır.

- **Random Forest**

Eğitim sırasında birçok karar ağacı oluşturarak çalışan sınıflandırma, regresyon için kullanılan topluluk öğrenme yöntemidir. Random Forest'lar karar ağaçların eğitim setine fazla uyma alışkanlığını düzeltir. Eğitim, makine öğrenene kadar bagging tekniğini uygular.

- **K-Neighbors**

K-Neighbors algoritması, uygulaması kolay gözetimli öğrenme algoritmalarındandır. Hem sınıflandırma hem de regresyon problemlerinin çözümünde kullanılır.

- **SVM**

SVM (Destek vektör makinesi), sınıflandırma ve regresyon problemleri için kullanılabilen denetimli bir makine öğrenmesi algoritmasıdır. Çoğunlukla sınıflandırma problemlerinde kullanılır. İki sınıftan iyi ayrım yapan hiper-düzlemi bularak sınıflandırma gerçekleştirir.

6. Test Sonuçları

Decision Tree kullanımı sonucu aldığım sonuçlar:

	precision	recall	f1-score	support
0	0.12	0.12	0.12	8
1	0.21	0.26	0.23	35
2	0.29	0.29	0.29	24
3	0.19	0.19	0.19	21
4	0.00	0.00	0.00	10
5	0.09	0.12	0.10	17
6	0.25	0.15	0.19	13
7	0.55	0.35	0.43	17
accuracy			0.21	145
macro avg	0.21	0.19	0.19	145
weighted avg	0.23	0.21	0.22	145

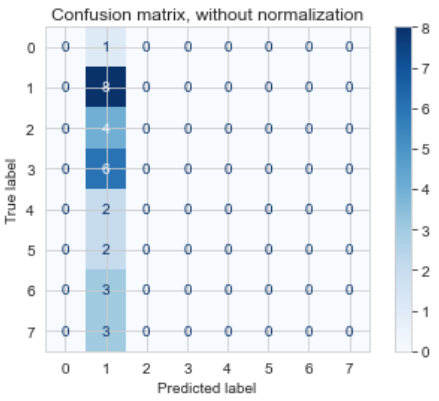
Random Forest kullanımı sonucu aldığım sonuçlar:

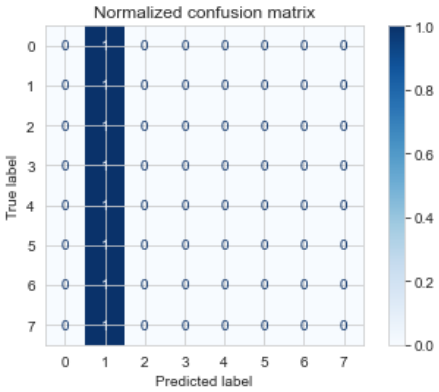
	precision	recall	f1-score	support
0	0.22	0.25	0.24	8
1	0.35	0.54	0.42	35
2	0.38	0.38	0.38	24
3	0.29	0.19	0.23	21
4	0.00	0.00	0.00	10
5	0.08	0.06	0.07	17
6	0.25	0.15	0.19	13
7	0.45	0.59	0.51	17
accuracy			0.32	145
macro avg	0.25	0.27	0.25	145
weighted avg	0.28	0.32	0.30	145

K-Neighbors kullanımı aldığım sonuçlar:

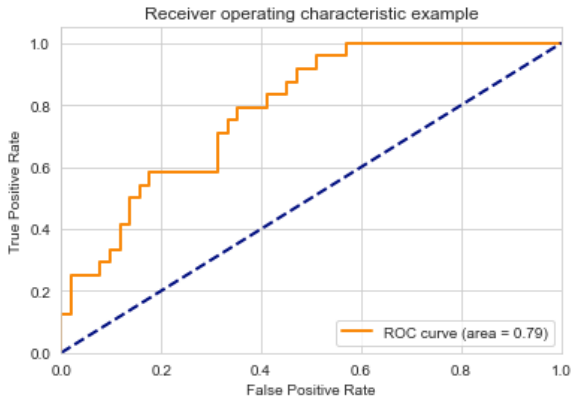
	precision	recall	f1-score	support
0	0.21	0.38	0.27	8
1	0.27	0.46	0.34	35
2	0.35	0.29	0.32	24
3	0.20	0.10	0.13	21
4	0.00	0.00	0.00	10
5	0.00	0.00	0.00	17
6	0.11	0.15	0.13	13
7	0.29	0.24	0.26	17
accuracy			0.23	145
macro avg	0.18	0.20	0.18	145
weighted avg	0.21	0.23	0.21	145

Modelin istatiksel hatalarını göstermek için Confusion Matrix'i kullanılmıştır.





Bu çalışma sonrası elde ettiğim bir Roc curve:



Yapılan işlemler sonrası ve çıkan sonuçlara göre Random Forest uygulaması diğer uygulamalara göre daha iyi bir sonuç vermiştir.

7. Sonuçlar (conclusions)

İlk olarak makine öğrenmesi algoritmaları karar ağacı ve rastgele orman karar ağacı eğitim veri seti üzerinde eğitilmiş ve modeller elde edilmiştir.
Veri seti %80 eğitim grubu ve %20 test grubu olarak ayrılmış olduğundan, ilk aşamada elde edilen eğitilmiş modeller test verilerine uygulanmıştır.
Her bir algoritmanın sonuçları, doğruluk, tahmin süresi ve hata oranı açısından analiz edilmiştir ve karşılaştırılmıştır.

Model	Accurarcy
Decision Tree	0.21
Random Forest	0.32
K-Neighbor	0.23

Elimdeki 145 öğrencinin girdilerini tutan veri seti ile bir makine öğrenmesi algoritması tasarlamaya çalıştım. Sırasıyla Decision Tree, Random Forest ve K-Neighbors uygulamalarını kullandım ve en iyi sonucu Random Forest sonrası aldım.

Bu çalışma sonucu derste işlenen bu uygulamaları kendim yapma şansı buldum ve böylece bu uygulamaların gerçekte nasıl yapıldığını ve ne tür sonuçlar alınabileceğini anladım.

Bu çalışmaların devamı olarak ileride kaggle'da bulunan yarışmalara katılmak isterim.

8. Referanslar

[1] https://scikit-learn.org/stable/auto_examples/model_selection/plot_confusion_matrix.html
[2] https://scikit-learn.org/stable/auto_examples/model_selection/plot_roc_crossval.html?highlight=roc+curve
[3] https://www.stackvidhya.com/plot-confusion-matrix-in-python-and-why/
[4] https://datatofish.com/confusion-matrix-python/
[5] https://www.kaggle.com/code/janietran/eda-on-higher-education-students-performance
[6] https://machinelearningmastery.com/regression-metrics-for-machine-learning/
[7] https://scikit-learn.org/stable/modules/tree.html
[8] https://ieeexplore.ieee.org/abstract/document/9077647
[9] https://medium.com/data-science-tr/sınıflandırma-modellerinde-başarı-kriterleri-2d86488799c6
[10] https://scikit-learn.org/stable/auto_examples/cluster/plot_cluster_iris.html#sphx-glr-auto-examples-cluster-plot-cluster-iris-py
[11] https://www.semanticscholar.org/paper/Student-Performance-Prediction-Model-using-Machine-Belachew-Gobena/d6940119e0aaf64d05e41484858808188567b7cf
[12] https://www.sciencedirect.com/science/article/pii/S1877050916300266