



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 becerililerin gelişmesini amaçlamaktadır. Staj süresince veri setinin toparlanması, veri setindeki kayıp değerlerin, hatalı veri girişlerin saptanması ve düzeltilmesi (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, Görselleştirme

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 Pre-processing), 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

1 Giriş	7
2 Bölüm 1	7
2.1 YÖK İZLEME VE DEĞERLENDİRME	7
2.2 Veri Setinin Tanıtımı	7
3 Bölüm 2	10
3.1 Yöntem	10
3.1.1 Veri Analizi Süreci	10
3.1.2 Web Scraper	11
3.1.3 Shapiro-Wilk Normallik Testi	12
3.1.4 K - Means	12
4 Bölüm 3	13
4.1 Power BI	13
4.2 Uygulama	14
4.2.1 Dönüşüm Adımı	16
4.2.2 Dummy uygulaması	17
4.2.3 Z standardizasyonu	18
4.2.4 K-Means	19
5 Sonuç	32
6 Kaynakça	33

Şekil Listesi

1	Veri Analiz Süreci	11
2	Power BI Panosu	14
3	Veri setinin tanımlayıcı bilgileri	15
4	Değişkenlere Ait Çarpıklık ve Basıklık Katsayıları	15
5	Değişkenlerin Shapiro-Wilk Testi	16
6	Dağılım Karşılaştırma	17
7	Dummy Ataması	17
8	Z-Score Normalizasyonu	18
9	K-küme Sayısı İçin Dirsek Yöntemi	19
10	Gözlemlerin küme ataması	20
11	Genel kümelemenin Silhouette skoru	21
12	Genel Kümeleme Grafiği	21
13	A grubu Elbow Yöntemi	22
14	A grubuna ait Scatter plot	23
15	A Grubunun Silhouette Değeri	23
16	B Grubu Elbow Yöntemi	24
17	B Grubu Kümeleme Grafiği	25
18	B Grubu Silhouette Skoru	26
19	C Grubu Elbow Yöntemi	27
20	C Grubunun Kümeleme Grafiği	28
21	C Grubu Silhouette Skoru	28
22	D Grubu Elbow Yöntem Grafiği	29
23	D Grubu Kümeleme Grafiği	30
24	D Grubu Silhouette Skoru	30

1 Giriş

2 Bölüm 1

2.1 YÖK İZLEME VE DEĞERLENDİRME

YÖK İzleme ve Değerlendirme Genel Raporu, üniversitelere ilişkin etraflı bir değerlendirme fırsatı sunan, yüksek öğretim sistemini iyileştirmeye yönelik ne tür müdahalelerde bulunabileceğine ve hangi alanda araştırmalar yapılabileceğine ilişkin ipuçları vereceği değerlendirmek amacı ile kullanılmaktadır.

2.2 Veri Setinin Tanıtımı

Üniversitelerin durumlarını belirlemek amacı ile 4 kategoriden oluşan ve toplamda 64 değişkeni olan ve üniversiteler tarafından yayınlanan raporların toparlanması neticesinde veri seti oluşturulmuştur. İçerisinde Dokuz Eylül Üniversitesinde bulunduğu bu veri setinde toplamda 200 üniversite bulunmaktadır. veri setinde bulunan değişkenler kodlanmıştır.

Değişkenlerin Kodları ve isimleri aşağıdaki gibidir;

- A.1: Mezun olan doktora öğrenci sayısı
- A.2.1: Kamu Personel Seçme Sınavlarında (KPSS) ilk % 5'lik dilime giren program sayısı
- A.2.2: Akademik Personel ve Lisansüstü Eğitimi Giriş Sınavlarında (ALES) ilk % 5'lik dilime giren program sayısı
- A.3: Uluslararası sempozyum, kongre veya sanatsal sergi sayısı
- A.4.1: Öğrencilerin yaptığı sosyal sorumluluk projelerinin sayısı
- A.4.2: Öğrencilerin yaptığı endüstriyel projelerin sayısı
- A.5: Teknokent veya Teknoloji Transfer Ofisi (TTO) projelerine katılan öğrenci sayısı
- A.6: Programların genel doluluk oranı
- A.7: Erişilebilen ders bilgi paketi oranı
- A.8: Mezun takip sistemi içerisindeki mezunların oranı
- A.9: Öğrencilerin kayıtlı oldukları program dışındaki diğer programlardan alabildikleri ders oranı

- A.10: Yükseköğretim Kurumları Sınavı (YKS) kılavuzunda akredite olduğu belirtilen lisans programı sayısı
- A.11.1: Üniversite kütüphanesinde öğrenci başına düşen basılı kitap sayısı
- A.11.2: Üniversite kütüphanesinde öğrenci başına düşen e-yayın sayısı
- B.1: Ulusal hakemli dergilerde yayımlanmış öğretim elemanı başına düşen yayın sayısı
- B.2: SCI, SCI-Expanded, SSCI ve AHCI endeksli dergilerde yayımlanmış öğretim elemanı başına düşen yayın sayısı
- B.3: En yüksek % 10'luk dilimde atıf alan yayın sayısı
- B.4: Üniversite adresli bilimsel yayınlara açık erişim oranı
- B.5.1: Başvurulan patent, faydalı model veya tasarım sayısı
- B.5.2: Sonuçlanan patent, faydalı model veya tasarım sayısı
- B.6: YÖK, TÜBA, TÜBİTAK bilim, teşvik ve sanat ödülleri sayısı
- B.7: YÖK 100/2000 Projesi doktora bursiyeri sayısı
- B.8: YÖK-YUDAB bursiyeri sayısı
- B.9: TÜBİTAK tarafından verilen ulusal ve uluslararası araştırma bursu sayısı
- B.10: TÜBİTAK tarafından verilen ulusal ve uluslararası destek programı sayısı
- B.11: Ulusal ve uluslararası özel veya resmi kurum ve kuruluşlar tarafından desteklenmiş Ar- Ge niteliği taşıyan proje sayısı
- B.12.1: Üniversitenin THE'ya göre dünya sıralaması
- B.12.2: Üniversitenin THE'ya göre bölgesel (Asya) sıralaması
- B.12.3: Üniversitenin THE'ya göre ulusal sıralaması
- B.12.4: Üniversitenin QS'e göre dünya sıralaması
- B.12.5: Üniversitenin QS'e göre bölgesel (Asya) sıralaması
- B.12.7: Üniversitenin ARWU'ya göre dünya sıralaması

- B.12.8: Üniversitenin ARWU'ya göre ulusal sıralaması
- B.13: Teknoloji Geliştirme Bölgelerinde (TGB) istihdam edilen doktora programlarına kayıtlı öğrenci sayısı
- B.14.1: Üniversite laboratuvarlarında Ar-Ge, inovasyon ve ürün geliştirme kapsamında sunulan hizmet sayısı
- B.14.2: Üniversite laboratuvarlarında Ar-Ge inovasyon ve ürün geliştirme kapsamında sunulan hizmetlerden elde edilen gelir
- B.15: Merkezi bütçe dışı öz gelir, döner sermaye, fon vb. gelirlerin yıllık bütçeye oranı
- B.16: Sağlık Uygulama ve Araştırma Merkezinin kâr ya da zararının toplam ciroya oranı
- B.17.1: Ar-Ge'ye harcanan bütçe oranı
- B.17.2: Ar-Ge'ye harcanan yatırım bütçesi oranı
- B.18.1: Endüstri ile ortak yürütülen proje sayısı
- B.18.2: Endüstri ile ortak yürütülen projelerin toplam bütçesi
- B.19: Yayın alımının bütçeye oranı
- C.1.1: İstihdam edilen yabancı uyruklu öğretim üyesi sayısı
- C.1.2: İstihdam edilen yabancı uyruklu doktoralı öğretim görevlisi ve araştırmacı sayısı
- C.2: Üniversitedeki yabancı uyruklu öğrenci sayısı
- C.3.1: Uluslararası değişim programları kapsamında gelen öğretim elemanı sayısı
- C.3.2: Uluslararası değişim programları kapsamında gönderilen öğretim elemanı sayısı
- C.4.1: Uluslararası değişim programları kapsamında gelen öğrenci sayısı
- C.4.2: Uluslararası değişim programları kapsamında gönderilen öğrenci sayısı
- C.5: Üniversite öğretim elemanlarının aldığı uluslararası fonlara dayalı proje sayısı

- C.6: Yurt dışındaki üniversiteler veya kurum ve kuruluşlar ile ortak yürütülen proje sayısı
- D.1: Üniversitenin yaptığı sosyal sorumluluk projesi sayısı
- D.2: Sürekli Eğitim Merkezi ve Dil Merkezi tarafından verilen sertifika sayısı
- D.3: Kariyer Merkezi çalışmaları kapsamında öğrenci ve mezunlara yönelik gerçekleştirilen faaliyet sayısı
- D.4: Kamu kurumları ile birlikte yürütülen proje sayısı
- D.5.1: Dezavantajlı gruplara yönelik sosyal entegrasyon ve kapsayıcılığa ilişkin yapılan faaliyet sayısı
- D.5.2: Üniversitenin engelsiz üniversite ödülü, engelsiz bayrak ödülü, engelsiz program nişanı ve engelli dostu ödülü sayısı
- D.6.1: Üniversitenin sıfır atık, yeşil kampüs ve çevrecilik alanlarında aldığı ödül sayısı
- D.6.2: Üniversitenin yeşil, çevreci üniversite endeksindeki sıralaması
- D.7: Üniversiteye kazandırılan bağış miktarı
- D.8: Öğrenci başına yapılan harcama miktarı
- D.9: Üniversitenin sağladığı eğitim burslarından faydalanan öğrenci oranı

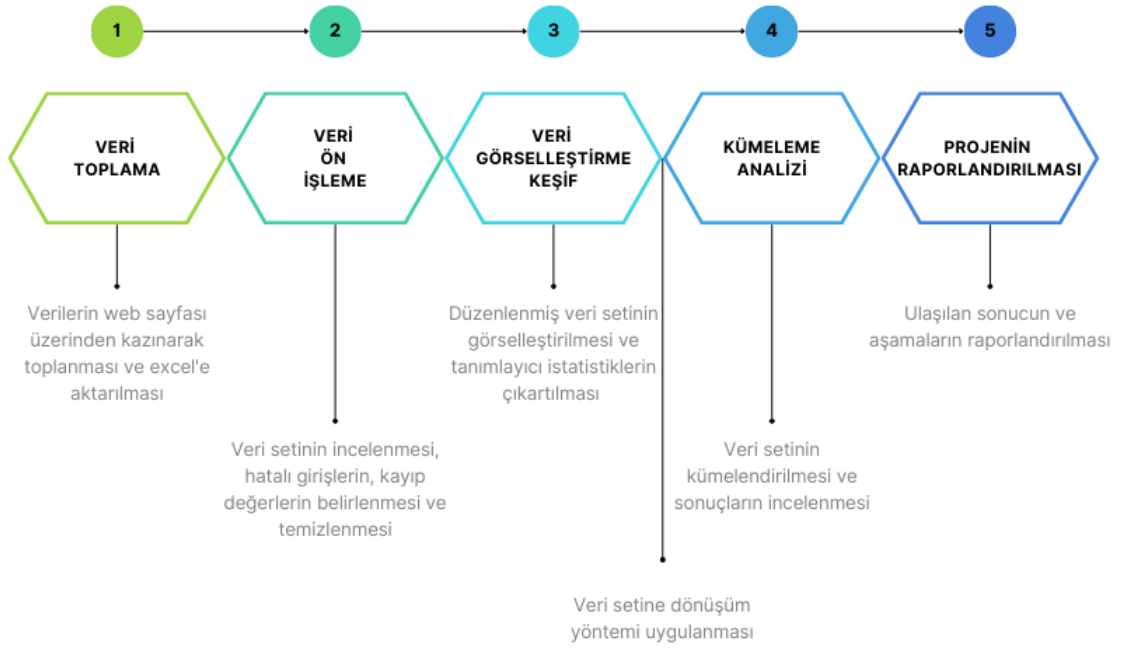
3 Bölüm 2

3.1 Yöntem

3.1.1 Veri Analizi Süreci

Yükseköğretim Kurulu'nun (YÖK) izleme ve değerlendirme raporları dikkate alınarak Dokuz Eylül Üniversitesi'nin grubunun kümeleme yöntemi kullanılarak belirlendiği ve süreç izleme modelinin tasarlandığı bir çalışmayı sunmaktadır. Üniversitenin performansını ve gelişimini izlemek, kaynakları etkin bir şekilde kullanmak ve karar verme süreçlerini iyileştirmek amacıyla oluşturulmuştur.

VERİ ANALİZ SÜRECİ



Şekil 1: Veri Analiz Süreci

3.1.2 Web Scraper

Web üzerinde çok büyük veriler tutulmaktadır fakat sorgulanması kolay değildir. Genelde veriler yapısal olmayan bir şekilde tutulur ve genellikle metin araması ile sorgulanır. Web içeriğinden bilgi çıkarılması için verinin yapısal bir biçime dönüştürülmesi gerekir. Bir çok kuruluş veri ihtiyaçlarını karşılayabilmek için web veri çıkarma yazılımları ortaya çıkmıştır. (Oğuz Kırat,2022,1-25)

Kullanılan veri setini internet üzerinden kazıyabilmek için Python'un bir kütüphanesi olan selenium kullanılmıştır. Selenium, farklı tarayıcılarda web m uygulama-

larını test etmek için kullanılan açık kaynaklı ve ücretsiz test aracıdır.

3.1.3 Shapiro-Wilk Normallik Testi

Shapiro-Wilk test bir veri setinin dağılımını kontrol eder ve normal dağılıma sahip olup olmadığını belirlemek için kullanılan bir hipotez testidir.

Shapiro-Wilk testin hipotezi şu şekildedir;

H0:Normal dağılır.

H1:Normal dağılmaz.

Shapiro-Wilk test istatistiği (W), gözlem değerlerinin sıralanması ve beklenen değerler arasındaki korelasyonu temel alır.

3.1.4 K - Means

K - Means algoritması makine öğrenim algoritmaları içerisinde en çok bilinen ve kullanılan algoritmalarından biridir. K-means algoritmasının genel mantığı n adet veri nesnesinden oluşan bir veri kümesini, giriş parametresi olarak verilen k adet kümeye bölümlenektir. Amacı ise gerçekleştirilen bölümlenme işlemi sonunda elde edilen kümelerin, küme içi benzerliklerinin maksimum ve kümeler arası benzerliklerinin minimum olmasını sağlamaktır.(SARIMAN, 2011)

K-Means algoritmasının işlem basamakları;

1. k adet nesne rastgele seçilir. Seçilen k adet nesne küme merkezlerini belirler.
2. Küme içi değişimler hesaplanır.
3. Her bir veriyi kendisine en yakın kümeye atar.
4. Verilerin hepsi en yakın kümelere atandığında tekrar k tane küme için merkezler hesaplanır.
5. Küme Merkezlerinde bir değişiklik olmayıncaya kadar 2. ve 3. Adımlar tekrarlanır.

K-means algoritmasının en büyük eksikliği k değerini tespit edememesidir. Bu nedenle başarılı bir kümeleme elde etmek için farklı k değerleri için denemeyanılma yönteminin uygulanması gerekmektedir. (DEMİRALAY ve ÇAMURCU,2005)

4 Bölüm 3

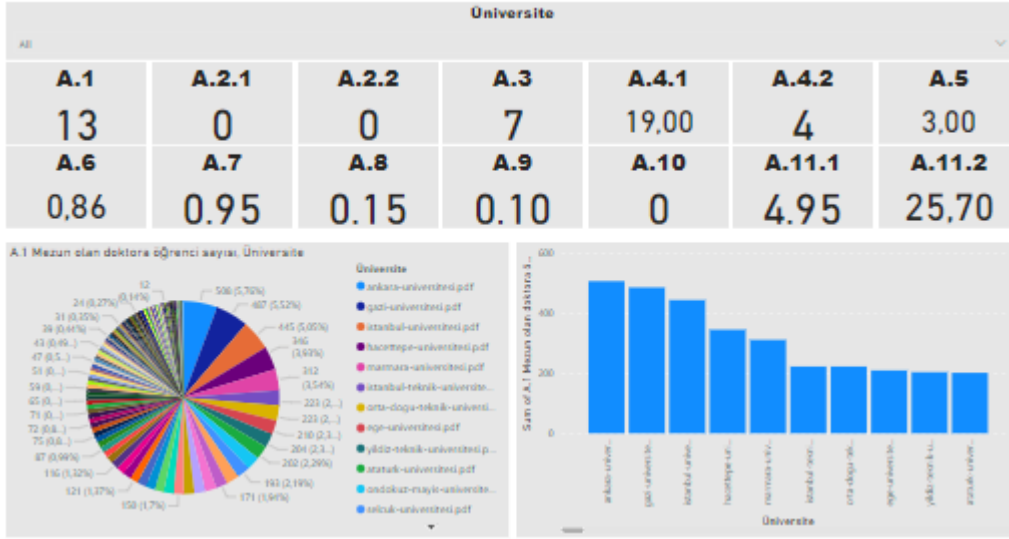
4.1 Power BI

Power BI, Microsoft tarafından geliştirilen bir iş zekası ve veri analitiği platformudur. Kullanıcılarına verileri görselleştirmek, analiz etmek ve paylaşmak için kullanabilecekleri bir dizi araç sunar. Power BI, çeşitli veri kaynaklarından veri alabilir, bunları birleştirebilir ve görsel raporlar, tablolar ve interaktif panolar oluşturabilir.

Veri setinin temel ön işleme uygulandıktan sonra veri setinin Power BI üzerinde nasıl düzenlendiği ve hazır hale getirildiği aşamalarına odaklanacağız. Power BI panosunu basit ve anlaşılır bir şekilde düzenlemek için aşağıdaki adımları izledik:

1. Veri setinin Ortama aktarılması
2. Veri setini anlama
3. Görselleştirme ekleme
4. Filtreleme ve sıralama
5. Panonun düzenlenmesi
6. Raporun export edilmesi

Sonuç olarak, temel ön işleme uygulanan veri setini Power BI üzerinde basit ve anlaşılır bir şekilde düzenledik. Panoda kullanıcı dostu görselleştirmeler ekledik, filtreleme ve sıralama işlemleri uyguladık ve raporun paylaşımını sağladık. Bu sayede, veri setindeki bilgilerin daha iyi anlaşılmasını ve analiz edilmesini sağladık.



Şekil 2: Power BI Panosu

4.2 Uygulama

Bu bölümde Yök izleme ve değerlendirme raporlarını yollamış üniversitelerin, bilgileri işlenmiş ve veri seti olarak kullanılmıştır. Gözlemler python'a ait selenium kütüphanesi ile yüksek öğretim kurumunun sayfası ziyaret edilerek elde edilmiştir. Kümele yöntemi olarak K-Means yöntemi kullanılmıştır. İlgili bölümde verilerin keşifsel analizi yapılmış ve sonuçlarına yer verilmiştir.

- Count(): Gözlem adedi
- Mean(): Değişkenin Ortalaması
- Std(): Değişkenin Standart Sapması
- Min(): Değişkenin İçerisindeki Minimum Değer
- Max(): Değişkenin İçerisindeki Maksimum Değer
- Skewness(): Değişkenin Çarpıklık Kat Sayısı
- Kurtosis(): Değişkenin Basıklık Kat Sayısı

	A.1	A.2.1	A.2.2	A.3	A.4.1	A.4.2	A.5	A.6	A.7	A.8	...
count	199.000000	199.000000	199.000000	199.000000	199.000000	199.000000	199.000000	199.000000	199.000000	199.000000	...
mean	44.296482	2.437186	2.693467	17.346734	60.870156	40.306533	65.413578	0.714633	0.642854	0.360553	...
std	80.258028	4.820497	5.410568	27.802249	124.603155	103.395736	138.241431	0.351270	0.444694	0.397264	...
min	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	...
25%	1.000000	0.000000	0.000000	3.000000	5.000000	0.000000	0.000000	0.731000	0.010000	0.008500	...
50%	13.000000	0.000000	0.000000	7.000000	19.000000	4.000000	3.000000	0.862000	0.950000	0.145000	...
75%	46.000000	3.000000	3.000000	19.000000	52.000000	28.500000	44.500000	0.946500	1.000000	0.790000	...
max	508.000000	27.000000	32.000000	204.000000	881.000000	894.000000	879.000000	1.000000	1.000000	1.000000	...

Şekil 3: Veri setinin tanımlayıcı bilgileri

Veri setindeki değişkenlerin dağılımları kontrol edilmiş ve ek olarak Shapiro-Wilk testi uygulanarak dağılımın normal dağılıp dağılmadığı gözlemlenmiştir.

	Çarpıklık Katsayısı	Basıklık Katsayısı
A.1	3.305418	13.044821
A.2.1	3.291121	11.959892
A.2.2	3.126392	10.548672
A.3	3.625549	16.913361
A.4.1	3.863438	17.103357
A.4.2	4.765203	28.517812
A.5	2.974397	9.866509
A.6	-1.369779	0.174844
A.7	-0.650957	-1.468215
A.8	0.629898	-1.299240
A.9	1.058698	0.285078
A.10	2.808029	8.706740
A.11.1	3.143063	11.248077
A.11.2	4.786016	25.011191
B.1	1.067668	5.342752
B.2	1.740613	6.496990
B.3	1.911254	3.760197
B.4	-0.518571	5.760385
B.5.1	3.479293	15.254582
B.5.2	3.664507	17.628859
B.6	5.435178	35.521509

Şekil 4: Değişkenlere Ait Çarpıklık ve Basıklık Katsayıları

Şekil 4'deki Çarpıklık ve Basıklık Katsayılarına bakıldığında veri setinin normal dağılmadığını gözlemleyebiliriz. Net bir kanıya varabilmek amacı ile Shapiro-wilk Testi uygulanır.

Shapiro-Wilk Testi için hipotezi;

H0: A.1 Değişkeni normal dağılır.

H1: A.1 Değişkeni normal dağılmaz.
Şeklinde kurulmalıdır.

	Shapiro-Wilk	-	P_Değeri
A.1	(0.581493616104126,	1.0537949808563125e-21)	
A.2.1	(0.5503978729248047,	2.134898432985204e-22)	
A.2.2	(0.5537185668945312,	2.521383983429303e-22)	
A.3	(0.5973562598228455,	2.4645123375775043e-21)	
A.4.1	(0.49870067834854126,	1.7949646374090276e-23)	
A.4.2	(0.4295768737792969,	8.76698299042269e-25)	
A.5	(0.5443150401115417,	1.5778479019691141e-22)	
A.6	(0.6839536428451538,	4.2352409919842214e-19)	
A.7	(0.6811925768852234,	3.539140072187292e-19)	
A.8	(0.7829029560089111,	6.758206953181309e-16)	
A.9	(0.8461697101593018,	3.0693790608118943e-13)	
A.10	(0.59396892786026,	2.0512592851151905e-21)	
A.11.1	(0.6365245580673218,	2.2520598021931388e-20)	
A.11.2	(0.437114417552948,	1.2012307450695356e-24)	
B.1	(0.9405854344367981,	2.734332724685373e-07)	
B.2	(0.8795076608657837,	1.6171059283309752e-11)	
B.3	(0.7767625451087952,	4.008126289787255e-16)	
B.4	(0.8621648550033569,	1.8948000720525604e-12)	
B.5.1	(0.5914729833602905,	1.793135111197639e-21)	
B.5.2	(0.5598121881484985,	3.430206524745167e-22)	
B.6	(0.3377014994621277,	2.3966991755381986e-26)	

Şekil 5: Değişkenlerin Shapiro-Wilk Testi

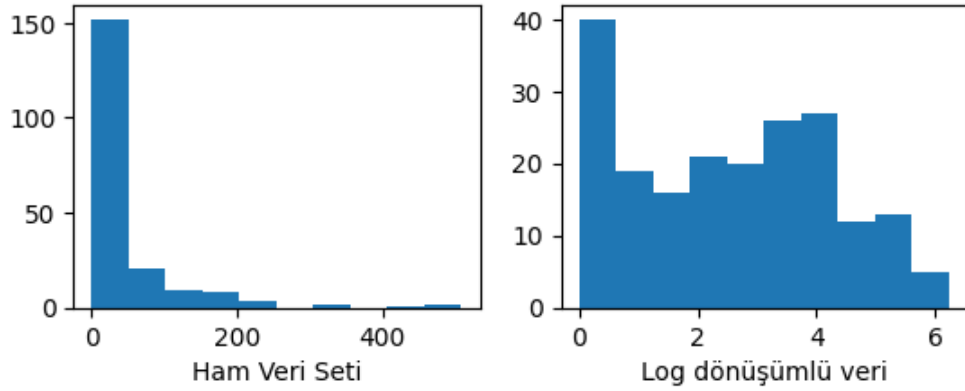
Şekil 5’de P-değeri belirlenmiş olan ($\alpha = 0.05$)’den küçük olduğu durumunda **H0** reddedilerek veri setinin normal dağılmadığı kanısına varılmıştır.

4.2.1 Dönüşüm Adımı

Veri setinin normal dağılmadığı tespit edilmiştir bu sebepten dolayı veri setinin istenilen çarpıklık ve basıklık kat sayısına ulaşmak için logaritmik dönüşüm işlemi uygulanmıştır.

Uygulanan logaritmik yöntem:

$$\log(x - \min + 1)$$



Şekil 6: Dağılım Karşılaştırma

4.2.2 Dummy uygulaması

Kategorik değişkenlerin analiz edilebilmesi için kategorik olarak tanımlanmış değişkenlerin kodlanması ile oluşturulan özel bir değişken türüdür. Veri setimde bulunan toplamda 7 adet kategorik değişkene dummy yöntemi ile atama yapıldı.

56	B.12.1_1001+	199	non-null	uint8
57	B.12.1_401-600	199	non-null	uint8
58	B.12.1_601-800	199	non-null	uint8
59	B.12.1_801-1000	199	non-null	uint8
60	B.12.2_1-200	199	non-null	uint8
61	B.12.2_201-400	199	non-null	uint8
62	B.12.2_401+	199	non-null	uint8
63	B.12.3_5	199	non-null	uint8
64	B.12.3_8	199	non-null	uint8
65	B.12.3_1-201	199	non-null	uint8
66	B.12.3_401+	199	non-null	uint8
67	B.12.4_1001+	199	non-null	uint8
68	B.12.4_401-600	199	non-null	uint8
69	B.12.4_601-800	199	non-null	uint8
70	B.12.4_801-1000	199	non-null	uint8
71	B.12.4_801-1001	199	non-null	uint8
72	B.12.5_1-200	199	non-null	uint8
73	B.12.5_201-400	199	non-null	uint8
74	B.12.7_401-600	199	non-null	uint8
75	B.12.7_601-800	199	non-null	uint8
76	B.12.7_801-1000	199	non-null	uint8
77	B.12.8_1_4	199	non-null	uint8
78	B.12.8_5_8	199	non-null	uint8

Şekil 7: Dummy Ataması

Şekil 7’de dummy yöntemi uygulandıktan sonra veri setine ek olarak 22 değişken daha eklenmiş oldu.

4.2.3 Z standardizasyonu

Z-skoru normalizasyonu, bu outlier sorununu önleyen verileri normalleştirme stratejisidir.

Z-skoru normalizasyonu için formül şu şekildedir:

$$z = \frac{x - \mu}{\sigma}$$

Veri setindeki tüm değişkenlere Standartlaştırma uygulandıktan sonra -1 ile 1 arasında yeni değerlerini almıştır.

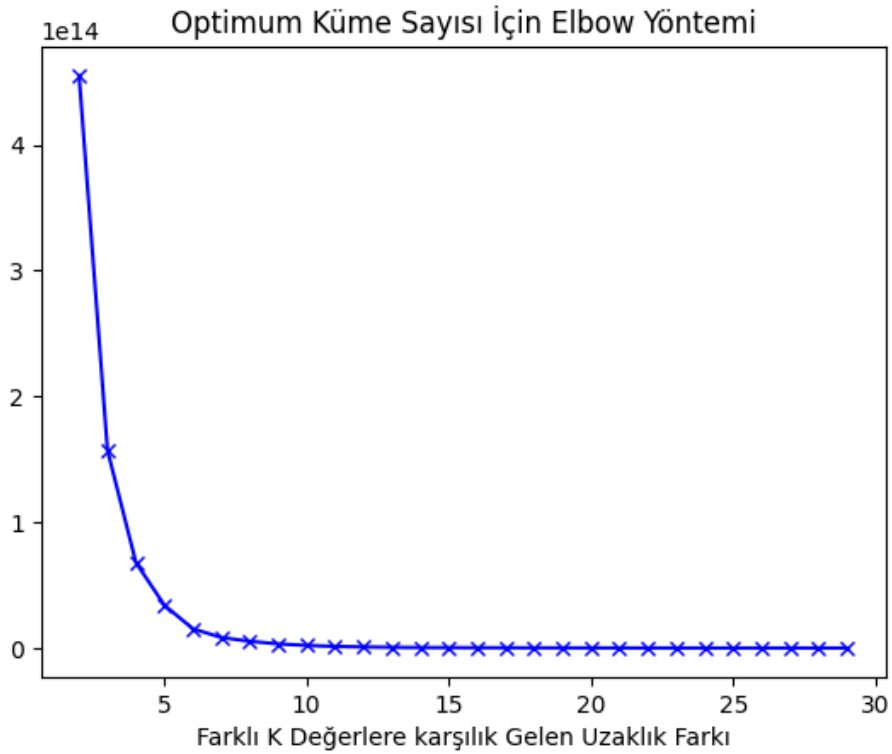
```
[-0.49656793, -0.80812002, 0.68859376, ..., -0.16054032,
-0.14322297, -0.14322297],
[-0.16613087, -0.80812002, -0.78468957, ..., -0.16054032,
-0.14322297, -0.14322297],
[-1.40134847, -0.80812002, -0.78468957, ..., -0.16054032,
-0.14322297, -0.14322297],
...,
[-0.30741275, -0.03377529, -0.0480479, ..., -0.16054032,
-0.14322297, -0.14322297],
[-1.40134847, -0.80812002, -0.78468957, ..., -0.16054032,
-0.14322297, -0.14322297],
[ 0.2827676, 0.74056943, 1.11950151, ..., -0.16054032,
-0.14322297, -0.14322297]]
```

Şekil 8: Z-Score Normalizasyonu

4.2.4 K-Means

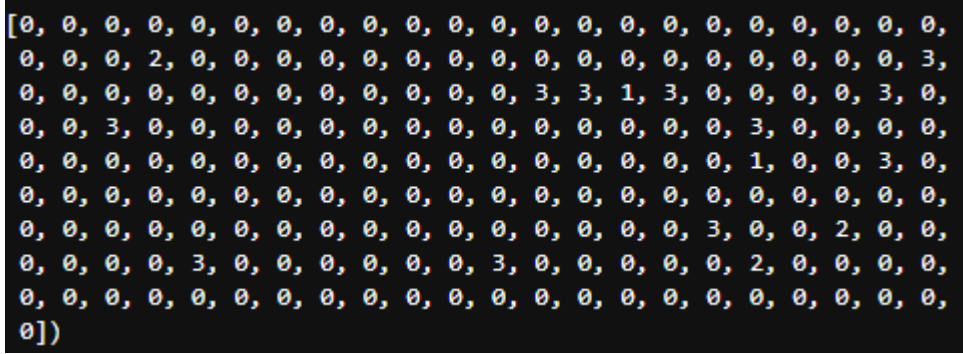
K-Means algoritması büyük çaptaki verisetleri üzerinde bile uygulaması kolay bir algoritmadır. Ancak algoritmanın basit ve verimli çalışmasını sağlayan bazı özellikleri aynı zamanda algoritma için dezavantajlarda yaratmaktadır. Parametre olarak verilen küme sayısı uygun olmayan k seçimlerinde kötü sonuçlar verebilir. (Taşkıran,2021)

Hatalı bir küme vermemek üzere 1 ile 30 arasında oluşacak şekilde bir küme oluşturuldu ve k kümesini bulabilmesi için elbow yöntemi (Dirsek Yöntemi) kullanıldı. Bu yöntem iki küme veya daha fazla kümeye ayrıldığında, her bir tekrar için bir ilave küme ekleyerek karesi alınmış mesafelerin toplamını maliyet fonksiyonu olarak hesaplar. Maliyet fonksiyonu değerleri, kümelerin alt sayısında keskin düşüş gösterebilir. Sonrasında bu düşüş, düzleşmeye başlar, bu düşüşün anlamı ilave kümeleme ile eklenen bilginin önceki kümeler sayısı kadar çok olmadığıdır. Dolayısıyla dirsek noktası olarak bilinen bu noktadaki kümelerin sayısı, kümelelerin optimal sayısı olarak seçilir.(Mohammed vd.,2018)



Şekil 9: K-küme Sayısı İçin Dirsek Yöntemi

Şekil 9'deki grafik göz önünde bulundurulduğunda veri setinde kullanılacak olan K küme sayısı grafikte en belirgin şekilde kırıldığı nokta en uygunudur. Kırılmanın fazla olduğu noktadan sonra hata da düşüşün yavaşlamaya başladığı noktadır, bu sebeb ile K=4 noktası küme sayısı en optimum sayıdır. K küme sayısı belirlendikten sonra kümeleme yapabilmek için veri setini fit ederek kümeleme için hazır hale getiriyoruz, ardından gözlemlerin küme merkezlerine olan uzaklıkları belirlenir ve bu uzaklıklara göre gözlemler kümelenir.



Şekil 10: Gözlemlerin küme ataması

Adımlar tamamlandıktan sonra Kümelenen verilerin bulunduğu kümedeki uygunluğunu bulmak için geliştirilen Silhouette yöntemi kullanıyoruz.

Silhouette hesaplamak için:

$$S = \frac{b - a}{\max(ab)}$$

a: bir örnek ile aynı kümedeki diğer tüm noktalar arasındaki ortalama mesafe.

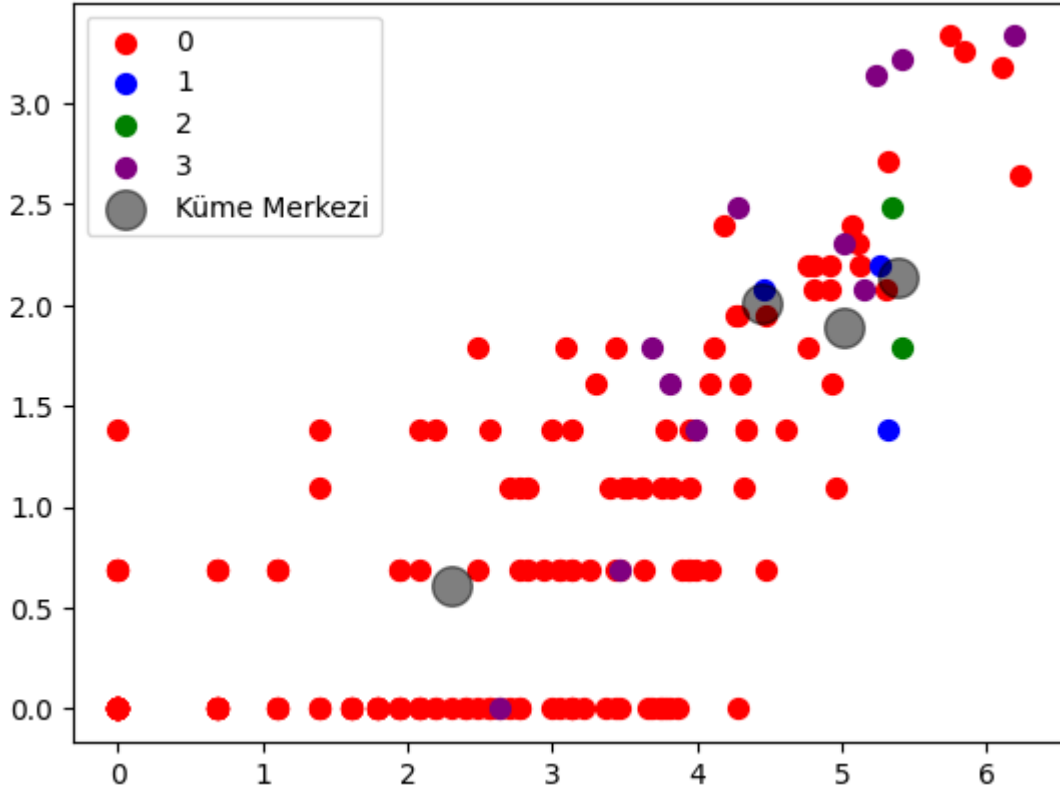
b: bir örnek ile en yakın kümedeki diğer tüm noktalar arasındaki ortalama mesafe.

Elde edilen Silhouette (S) değeri -1 ile 1 arasında bir sayıdır. Bu değer 1'e yakın ise kümelemenin iyi olduğu anlamına gelmektedir. (Aslanyürek ve Mesut,2021)

```
silhouette_score  
  
0.8991081593745532
```

Şekil 11: Genel kümelemenin Silhouette skoru

Şekil 11’de görüldüğü üzere K-Means yönteminin veri setini ne kadar doğru bir şekilde kümelediğini ve genel kümeleme analizinde veri setinin yoğunluk sağlama düzeyini değerlendirmek amacıyla Silhouette yöntemi kullanılmış ve S değeri elde edilmiştir. Sonuçlara göre K-Means yönteminin veri setini 0.8991 oranında doğru bir şekilde kümelediğini ve genel kümeleme analizinde veri setinin yoğunluk sağlama düzeyi belirlenmiştir.



Şekil 12: Genel Kümeleme Grafiği

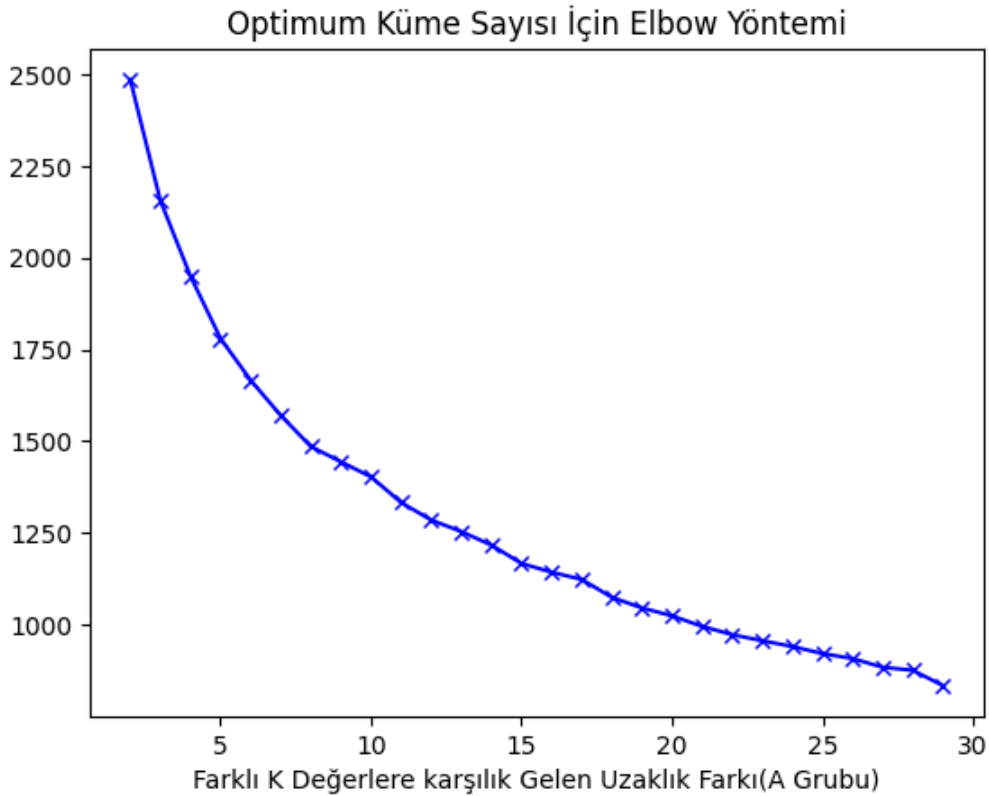
Şekil 12’deki grafikte 4 kümeye ayrılan veriler gözlemlenmektedir. 0 grubu çoğunlukta olup daha geniş alana yayılmışken 3 grubu dağınık dağılmış bir şekilde gözlemlenmektedir. Tüm gruplardaki gözlemler birbirine benzer özelliğe sahip olan kümeler ile eşleştirilmiştir.

Değişkenlerin Grupsal Kümelenmesi

A Grubu Kümeleme

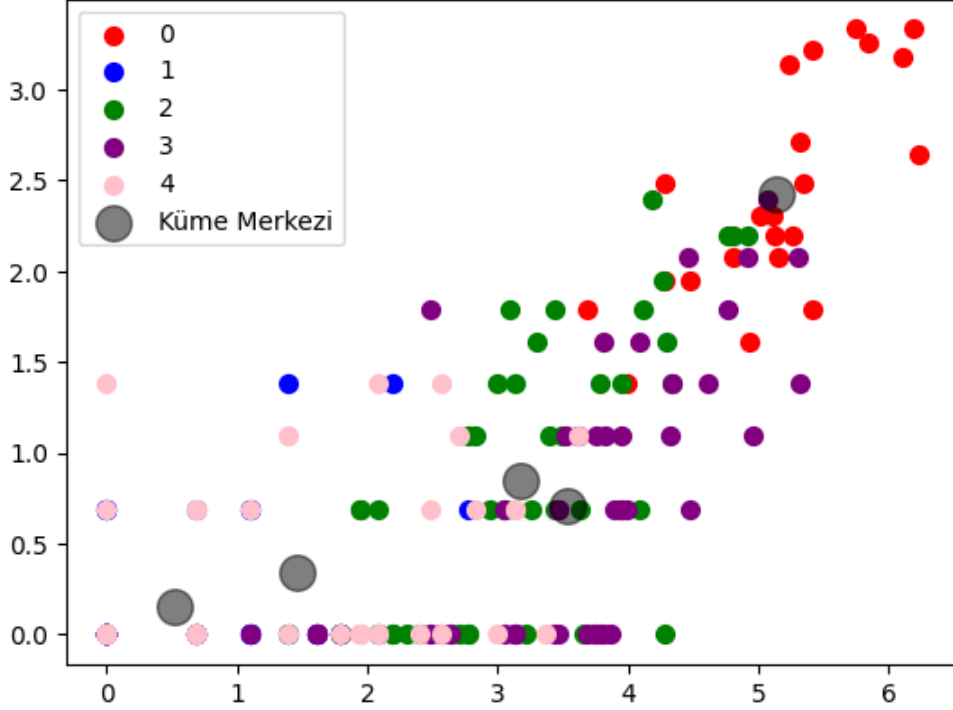
Veri setindeki değişkenler kendi içlerinde 4 grup olarak belirlenmiştir, bu grupların kümelede hangisinin daha baskın olduğu merak edilmiş ve gruplar arasında kümeleme yapılmıştır.

K-Means yöntemi kullanılarak A grubu üzerinde gerçekleştirilen genel kümeleme analizi için k küme sayısı sonuçlarına göre k küme sayısı belirlenmiştir. Aşağıdaki grafikte bu sonuçlar görülebilir:



Şekil 13: A grubu Elbow Yöntemi

Elbow yöntemi grafi (Şekil 13) sonuçları değerlendirildiğinde küme sayısının K=5 olması daha uygun olduğu kanısına verilmiştir. K=5 küme sayısına göre K-Means uygulanmıştır.



Şekil 14: A grubuna ait Scatter plot

K-Means yönteminin A grubunu ne kadar doğru bir şekilde kümelediği ve genel kümeleme analizinde A grubunun yoğunluk sağlama düzeyini değerlendirmek amacıyla Silhouette yöntemi kullanılmış ve S değeri elde edilmiştir. Sonuçlara göre K-Means yönteminin A grubunu ne kadar doğru bir şekilde kümelediğini ve genel kümeleme analizinde A grubunun yoğunluk sağlama düzeyini belirlemek için Silhouette yönteminin etkili bir değerlendirme aracı olduğunu gösterecektir.

```
silhouette_score_2  
0.16418174418364956
```

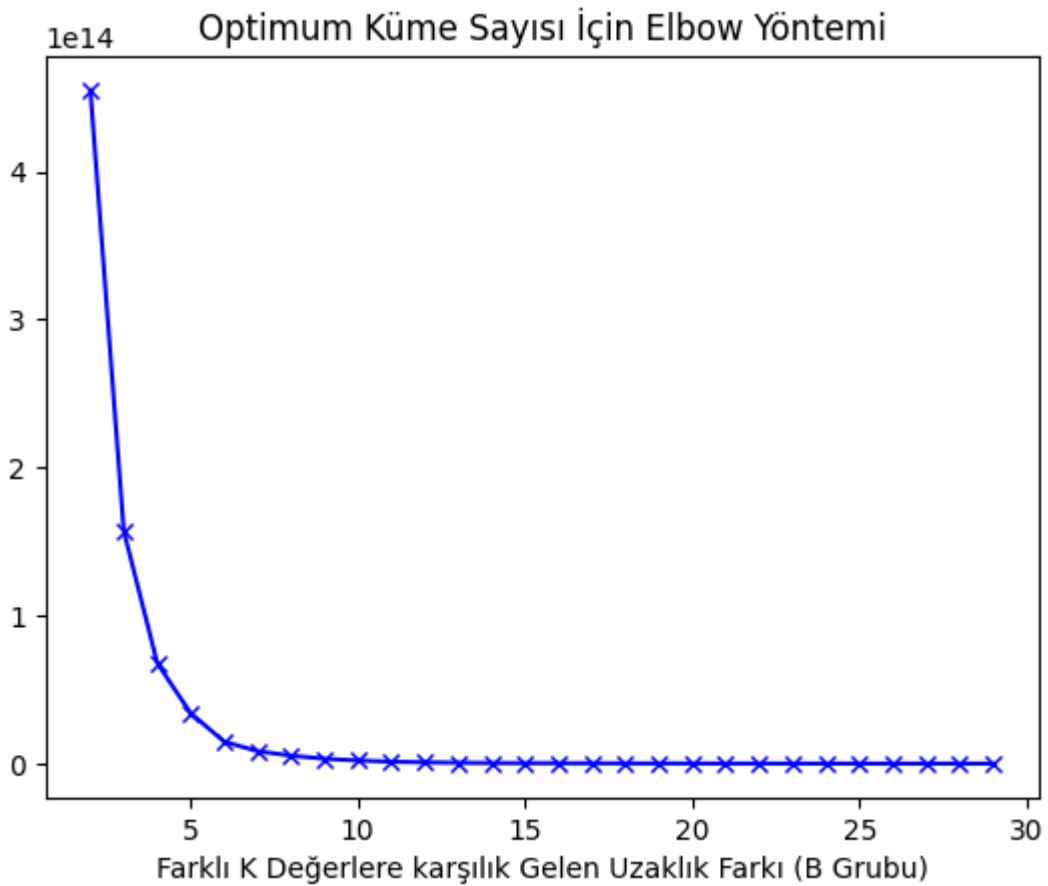
Şekil 15: A Grubunun Silhouette Değeri

Silhouette yöntemi ile belirlenen S değeri 0,1614 olarak belirlenmiştir. K-ortalamlar

yöntemiyle yapılan kümeleme analizinde A grubunun genişleme yoğunluğunun düşük olduğunu vurgulanmıştır. Bu S değeri, A grubunun diğer kümelerle göre daha az benzerlik gösterdiği görülmektedir. K-ortalamalarının A grubunu doğru bir şekilde gruplandırmasının zor olduğunu ve bu grubun diğer verilerden farklı özelliklere sahip olduğu görülmektedir. A grubunun sonuçları göz önünde bulundurularak Dokuz Eylül Üniversitesi'nin grubu belirlenmiş, belirlenen grubun içerisinde Başkent Üniversitesi, Boğaziçi Üniversitesi, Yıldız Teknik Üniversitesi gibi 34 üniversite bulunmaktadır.

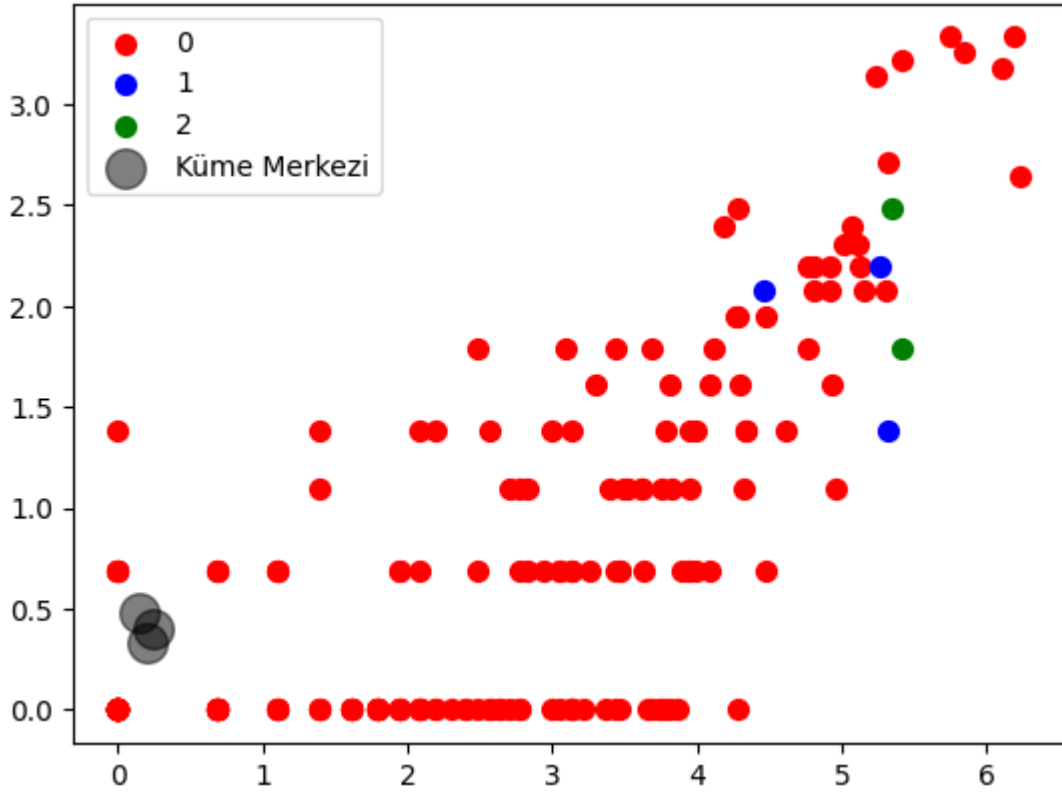
B Grubu Kümeleme

K-Means yöntemi kullanılarak B grubu üzerinde gerçekleştirilen genel kümeleme analizi için k küme sayısı sonuçlarına göre k küme sayısı belirlenmiştir. Aşağıdaki grafikte bu sonuçlar görülebilir:



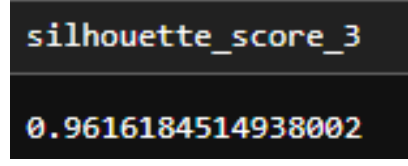
Şekil 16: B Grubu Elbow Yöntemi

Şekil 16'daki grafik incelendiğinde, farklı k değerlerine karşılık gelen kümeleme sonuçlarının performansını değerlendirebiliriz. Değerlendirme sonuçlarına dayanarak, en uygun k değeri olarak 3 değeri seçildi. Seçilen k değeriyle gerçekleştirilen kümeleme analizinin detayları ve grafiğine aşağıda yer verilmiştir.



Şekil 17: B Grubu Kümeleme Grafiği

Veri setinin B grubu üzerinde K-Means yöntemi kullanılarak yapılan kümeleme sonuçları gösterilmektedir. Gözlemlediğimiz sonuçlara göre, her bir kümenin merkezi, grafiğin üzerinde koyu renkle işaretlenmiştir. Bu merkez noktaları, ilgili kümeye ait veri noktalarının genel yoğunluğunu temsil etmektedir. K-Means algoritması tarafından oluşturulan bu kümeleme sonucu, B grubu veri noktalarının belirli desenler veya benzerlikler gösterdiği görülmektedir.

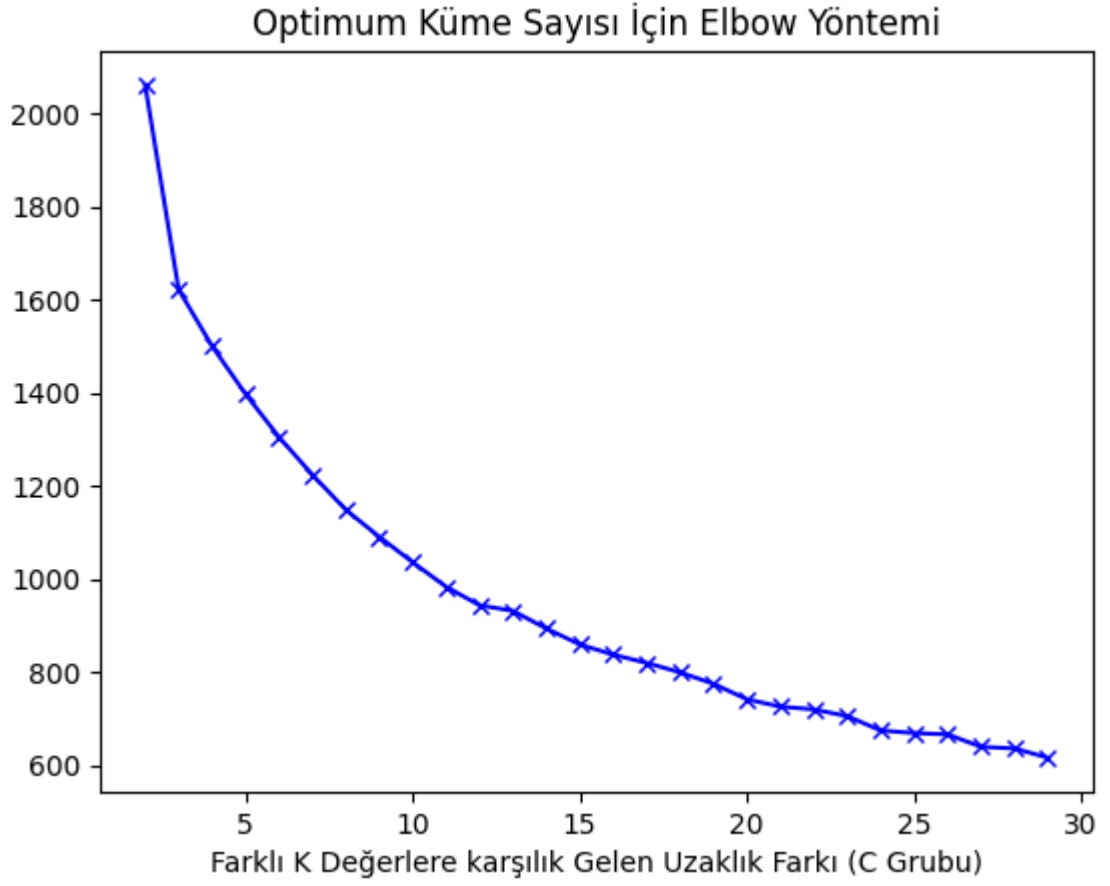
A black rectangular box with white text. The top line reads 'silhouette_score_3' and the bottom line reads '0.9616184514938002'.

Şekil 18: B Grubu Silhouette Skoru

Silhouette yöntemi ile belirlenen S değeri 0.9616 olarak belirlenmiştir. K-ortalamlar yöntemiyle yapılan kümeleme analizinde B grubunun genişleme yoğunluğunun yüksek olduğu vurgulanmıştır. Bu S değeri, B grubunun diğer kümelerle göre daha fazla benzerlik gösterdiği görülmektedir. K-ortalamlarının B grubunu doğru bir şekilde gruplandırmasının kolay olduğunu ve bu grubun diğer verilerden farklı özelliklere sahip olduğu görülmektedir. B grubunun sonuçları göz önünde bulundurularak Dokuz Eylül Üniversitesinin grubu belirlenmiş, belirlenen grubun içerisinde Ankara Üniversitesi, Boğaziçi Üniversitesi, Eskişehir Osmangazi Üniversitesi gibi 194 üniversite bulunmaktadır.

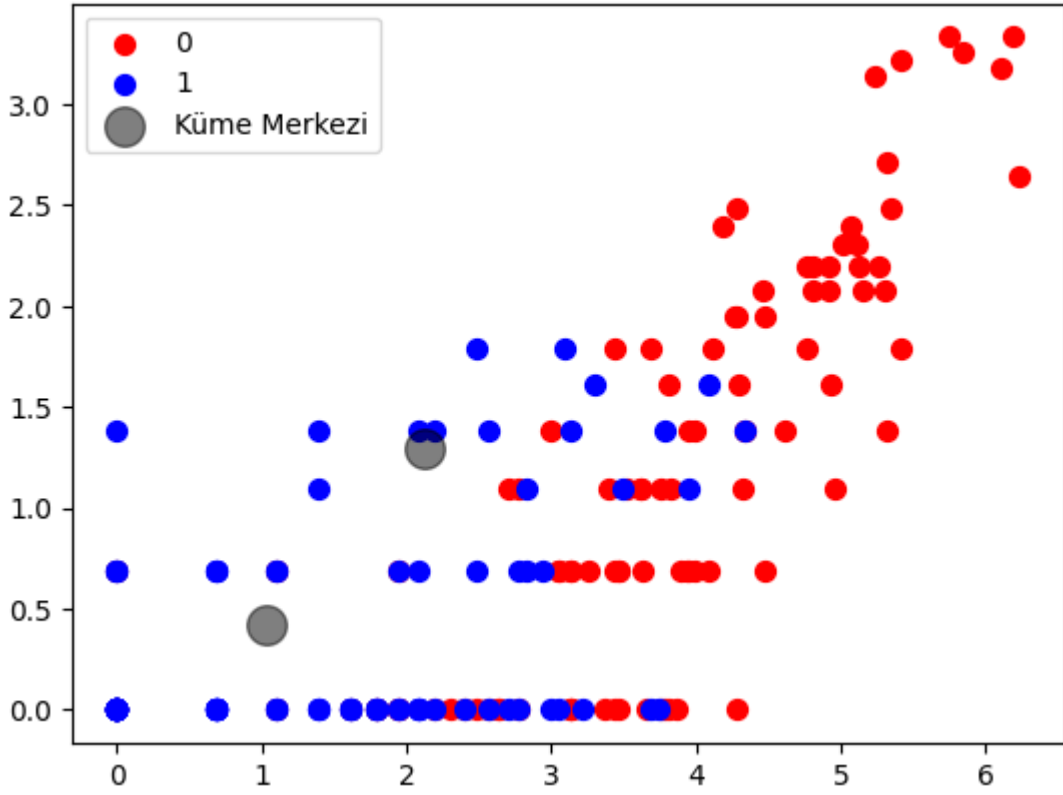
C Grubu Kümeleme

K-Means yöntemi kullanılarak C grubu üzerinde gerçekleştirilen genel kümeleme analizi için k küme sayısı sonuçlarına göre k küme sayısı belirlenmiştir. Aşağıdaki grafikte bu sonuçlar görülebilir:



Şekil 19: C Grubu Elbow Yöntemi

Şekil 19'daki grafik incelendiğinde, farklı k değerlerine karşılık gelen kümeleme sonuçlarının performansını değerlendirebiliriz. Değerlendirme sonuçlarına dayanarak, en uygun k değeri olarak 2 değeri seçildi. Seçilen k değeriyle gerçekleştirilen kümeleme analizinin detayları ve grafiğine aşağıda yer verilmiştir.



Şekil 20: C Grubu Kümeleme Grafiği

Veri setinin C grubu üzerinde K-Means yöntemi kullanılarak yapılan kümeleme sonuçları gösterilmektedir. Gözlemlediğimiz sonuçlara göre, her bir kümenin merkezi, grafiğin üzerinde koyu renkle işaretlenmiştir. Bu merkez noktaları, ilgili kümeye ait veri noktalarının genel yoğunluğunu temsil etmektedir. K-Means algoritması tarafından oluşturulan bu kümeleme sonucu, C grubu veri noktalarının belirli desenler veya benzerlikler gösterdiği görülmektedir.

```
silhouette_score_4  
0.27868474890130185
```

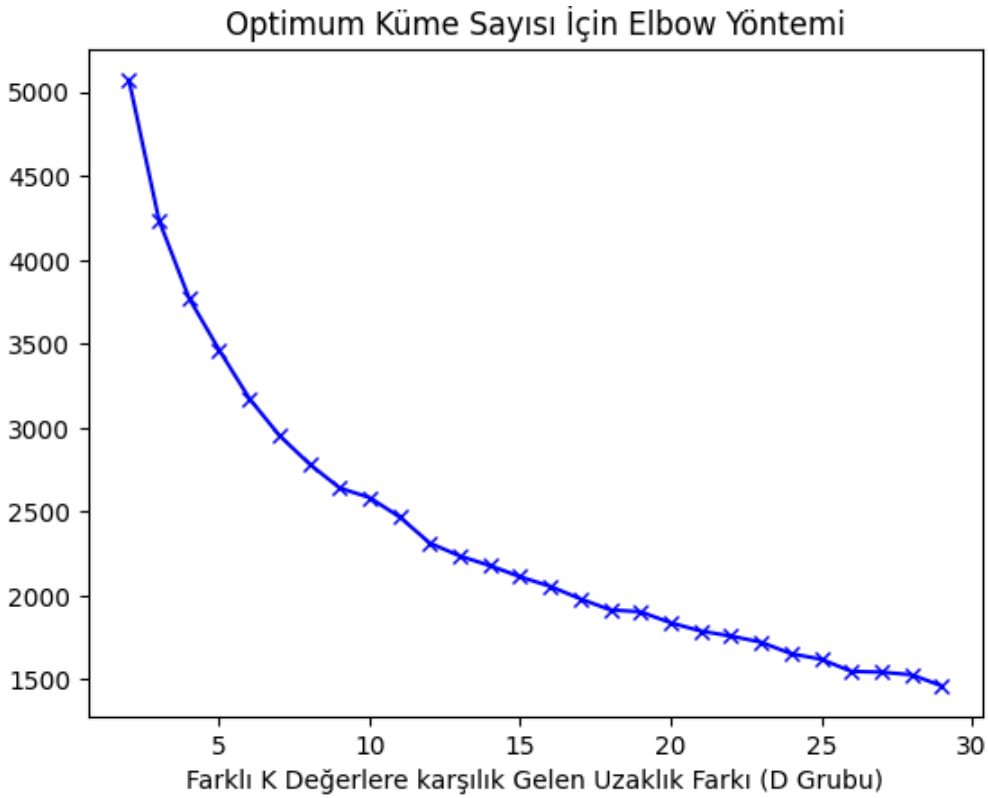
Şekil 21: C Grubu Silhouette Skoru

Silhouette yöntemi ile belirlenen S değeri 0.2786 olarak belirlenmiştir. K-ortalamlar yöntemiyle yapılan kümeleme analizinde C grubunun genişleme yoğunluğunun düşük olduğunu vurgulanmıştır. Bu S değeri, C grubunun diğer kümelerle göre

daha az benzerlik gösterdiği görülmektedir. K-ortalamalarının C grubunu doğru bir şekilde gruplandırmasının zor olduğunu ve bu grubun diğer verilerden farklı özelliklere sahip olduğu görülmektedir. C grubunun sonuçları göz önünde bulundurularak Dokuz Eylül Üniversitesinin grubu belirlenmiş, belirlenen grubun içerisinde Ege Üniversitesi, Düzce Üniversitesi, Gazi Üniversitesinin de bulunduğu bu grupta toplamda 94 üniversite bulunmaktadır.

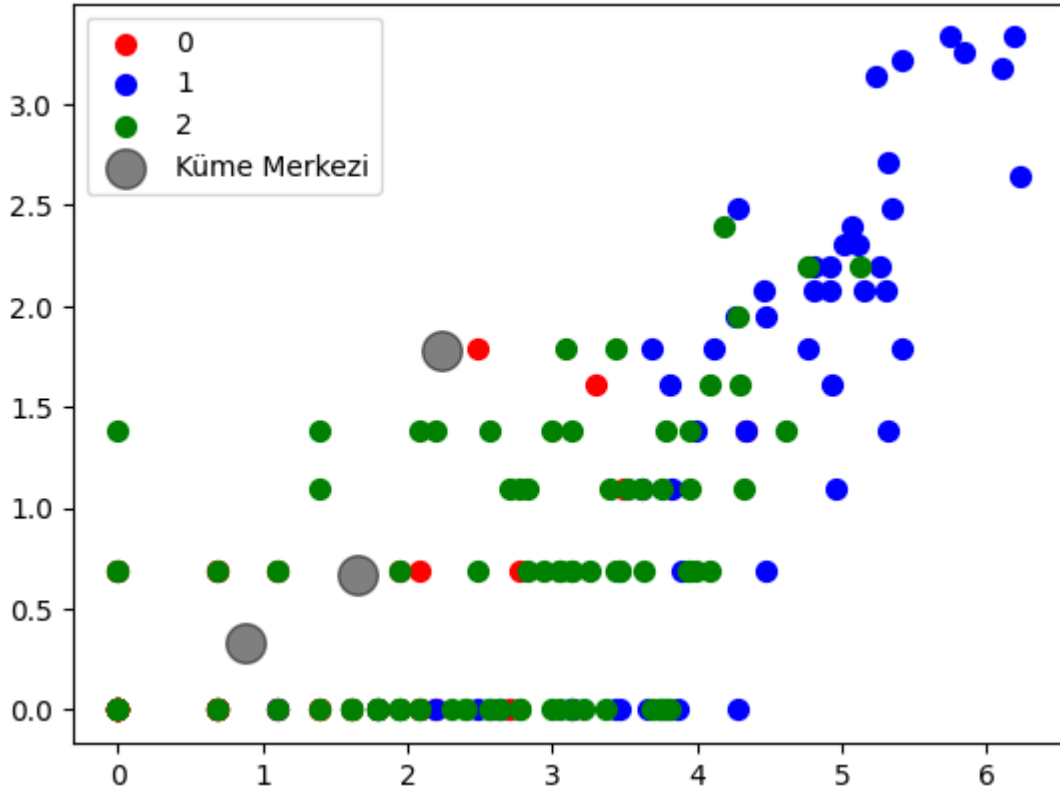
D Grubu Kümeleme

K-Means yöntemi kullanılarak C grubu üzerinde gerçekleştirilen genel kümeleme analizi için k küme sayısı sonuçlarına göre k küme sayısı belirlenmiştir. Aşağıdaki grafikte bu sonuçlar görülebilir:



Şekil 22: D Grubu Elbow Yöntem Grafiği

Şekil 22'deki grafik incelendiğinde, farklı k değerlerine karşılık gelen kümeleme sonuçlarının performansını değerlendirebiliriz. Değerlendirme sonuçlarına dayanarak, en uygun k değeri olarak 3 değeri seçildi. Seçilen k değeriyle gerçekleştirilen kümeleme analizinin detayları ve grafiğine aşağıda yer verilmiştir.



Şekil 23: D Grubu Kümeleme Grafiği

Veri setinin D grubu üzerinde K-Means yöntemi kullanılarak yapılan kümeleme sonuçları gösterilmektedir. Gözlemlediğimiz sonuçlara göre, her bir kümenin merkezi, grafiğin üzerinde koyu renkle işaretlenmiştir. Bu merkez noktaları, ilgili kümeye ait veri noktalarının genel yoğunluğunu temsil etmektedir. K-Means algoritması tarafından oluşturulan bu kümeleme sonucu, D grubu veri noktalarının belirli desenler veya benzerlikler gösterdiği görülmektedir.

```
silhouette_score_5
-0.013085471133520014
```

Şekil 24: D Grubu Silhouette Skoru

Silhouette yöntemi ile belirlenen S değeri -0.0130 olarak belirlenmiştir. K-ortalamalar yöntemiyle yapılan kümeleme analizinde D grubunun genişleme yoğunluğunun

çok düşük olduđuunu vurgulanmıřtır. Bu S deęeri, D grubunun dięer k melerle g re  ok az benzerlik g sterdięi g r lmektedir. K-ortalamlarının D grubunu doęru bir řekilde gruplandırmasının zor olduđuunu ve bu grubun dięer verilerden farklı  zelliklere sahip olduđu g r lmektedir. D grubunun sonu ları g z  n nde bulunduurularak Dokuz Eyl l  niversitesinin grubu belirlenmiř, belirlenen gurubun i erisinde Bařkent  niversitesi, Boęazi i  niversitesi, Ege  niversitesi gibi 50  niversite bulunmaktadır.

5 Sonuç

Bu çalışmada, veri setinin genel kümeleme yöntemi olan K-Means kullanılarak bir analiz gerçekleştirildi. Elde edilen sonuçlar, veri setine ait noktaların kümeleme yapılabilir bir yapıya sahip olduğunu ortaya koydu. K-Means algoritmasıyla oluşturulan grafikler, veri noktalarının belirli desenler veya benzerlikler gösterdiğini göstermektedir.

K-Means algoritmasının elde ettiği sonuçlar, Değişken alt gruplarındaki veri noktalarının belirli özellikler veya ilişkiler temelinde gruplandığını göstermektedir. Değişken gruplarından biri olan B grubunun içerdiği verilerin benzerliklerine ve farklılıklarına bakılarak B grubunun daha baskın bir kümeleme sahip olduğunu göstermiştir.

Dokuz Eylül Üniversitesi, yapılan kümeleme analizi sonucunda Boğaziçi Üniversitesi, Orta Doğu Teknik Üniversitesi gibi benzer üniversitelerle aynı küme içerisinde yer almaktadır. Bu sonuçlar, Dokuz Eylül Üniversitesi'nin stratejik planlamalarında B grubu değişkenlerini dikkate alması gerektiğini öne çıkarmaktadır. B grubu değişkenleri analiz edilerek Dokuz Eylül Üniversitesi'nin güçlü yönleri, zayıf yönleri, fırsatları belirlenebilir. Bu analiz, üniversitenin mevcut durumunu değerlendirmesine, potansiyel stratejik hedefler belirlemesine ve kaynaklarını etkili bir şekilde yönlendirmesine yardımcı olabilir.

Sonuç olarak, Dokuz Eylül Üniversitesi'nin kümeleme analizi sonuçlarına dayanarak B grubu değişkenlerini dikkate alması, stratejik planlamalarını güçlendirmek ve benzer üniversitelerle rekabet edebilme kapasitesini artırmak için önemli bir adımdır. Bu analiz, üniversitenin gelecekteki gelişim stratejilerinin oluşturulmasına ve başarılı bir şekilde uygulanmasına katkı sağlayabilir.

6 Kaynakça

SARIMAN,G.,(2011),K-Means ve K-Medoids Kümeleme Algoritmalarının Karşılaştırılması, *Süleyman Demirel Üniversitesi Fen Bilimleri Dergisi*,192-202

DEMİRALAY,M.,ÇAMURCU,A.,Y.,(2005)KÜMELEME YETENEKLERİNİN KARŞILAŞTIRILMASI,*İstanbul Ticaret Üniversitesi Fen Bilimleri Dergisi*,8,1-18

KIRAT,O.,(2022),*Ajax Tabanlı Web Sayfalarından Veri Çıkarımına Bir Yaklaşım*,Yüksek Lisans Tezi,TRAKYA ÜNİVERSİTESİ

TAŞKIRAN,S.F.,(2022),*Doğal Dil İşleme ile Akademik Metinlerin Kümelenmesi*,Yüksek Lisans Tezi,Konya Teknik Üniversitesi

Mohammed, Z.,AL-SHEHABI,S.,Dökeroğlu,T.,(2018),Gözetimsiz Makine Öğrenme Teknikleri ile Miktarla Dayalı Negatif Birliktelik Kural Madenciliği,*Bilim ve Teknoloji Dergisi*, 6, 1119-1138

Aslanyürek, M.,Mesut, A.,(2021), Kümeleme Performansını Ölçmek için Yeni Bir Yöntem ve Metin Kümeleme için Değerlendirmesi, *Avrupa Bilim ve Teknoloji Dergisi*,27, 53-65