T.C.
YILDIZ TEKNİK ÜNİVERSİTESİ

ELEKTRİK-ELEKTRONİK FAKÜLTESİ

BİLGİSAYAR MÜHENDİSLİĞİ BÖLÜMÜ

Yapay Zeka

2. ÖDEV

Ödevin Konusu
Bir veri kümesi oluşturup makine öğrenmesi algoritmalarını çalıştırmak
(Lineer Regresyon ile Ev Kirası Tahmin Etme)

Dersi Veren Öğretim Üyesi
Doç. Dr. Mehmet Fatih AMASYALI

Ödevi Yapan Öğrenciler

16014023	Eray Zeki SAH
16014033	Ferhat TAŞ

Ödev Teslim Tarihi
02.06.2021

İÇİNDEKİLER

KONU SEÇİMİ VE VERİ KÜMESİ OLUŞTURMA

1.1 Konu Seçimi

Ödevde konu; “Lineer Regresyon ile İstanbul ili Beşiktaş ilçesinde bulunan dairelerin aylık kira tahmini” olarak belirlenmiştir. Konu kapsamında kullanılacak veri kümesi internet ortamından çekilmiştir. Veri kümesi çekilirken evlere ait aşağıdaki özellikler alınmış ve lineer regresyonda kullanılmıştır:

- Kira
- Alan
- Oda Sayısı
- Banyo Sayısı
- WC Sayısı
- Bina Yaşı
- Isıtma Tipi
- Balkon Durumu
- Eşya Durumu
- Site İçerisinde

1.2 Veri Seti İndirme

Veri kümesi, Türkiye’de en çok kullanılan kiralık ev bulma web sitelerinden birisi olan https://www.emlakjet.com/ üzerinden bir crawler yardımıyla çekilmiştir.

Buna göre, konu seçimi aşamasında belirtilmiş olan her bir eve ait bilgiler çekilerek csv formatında text olarak kaydedilmiştir.

Toplamda, bazı evlerde “belirtilmemiş” olan kısımlar bulunduğundan o evler çıkartılarak 512 adet veri elde edilmiştir.

1.3 Veri Seti Düzenleme (Data Pre-processing)

Veri kümesinde; “Oda Sayısı”, “Bina Yaşı”, “Balkon Durumu”, “Eşya Durumu”, “Site İçerisinde” ve “Isıtma Tipi” verileri, sayısal olarak değil kategorisel olarak bulunmaktadır. Bu veriler üzerinde Linear Regresyon yapılabilmesi için sayısal olarak ifade edilmeleri gerekmektedir. Bu probleme çözüm üretebilmek adına, “Isıtma Tipi” haricindeki tüm veriler, sayısal karşılıklarıyla bir map vasıtasıyla değiştirilmiştir. Sayısal olarak ifade edilmesi daha zor olan “Isıtma Tipi” verisi ise Dummy Coding yöntemi ile sayısal hale getirilmiştir.

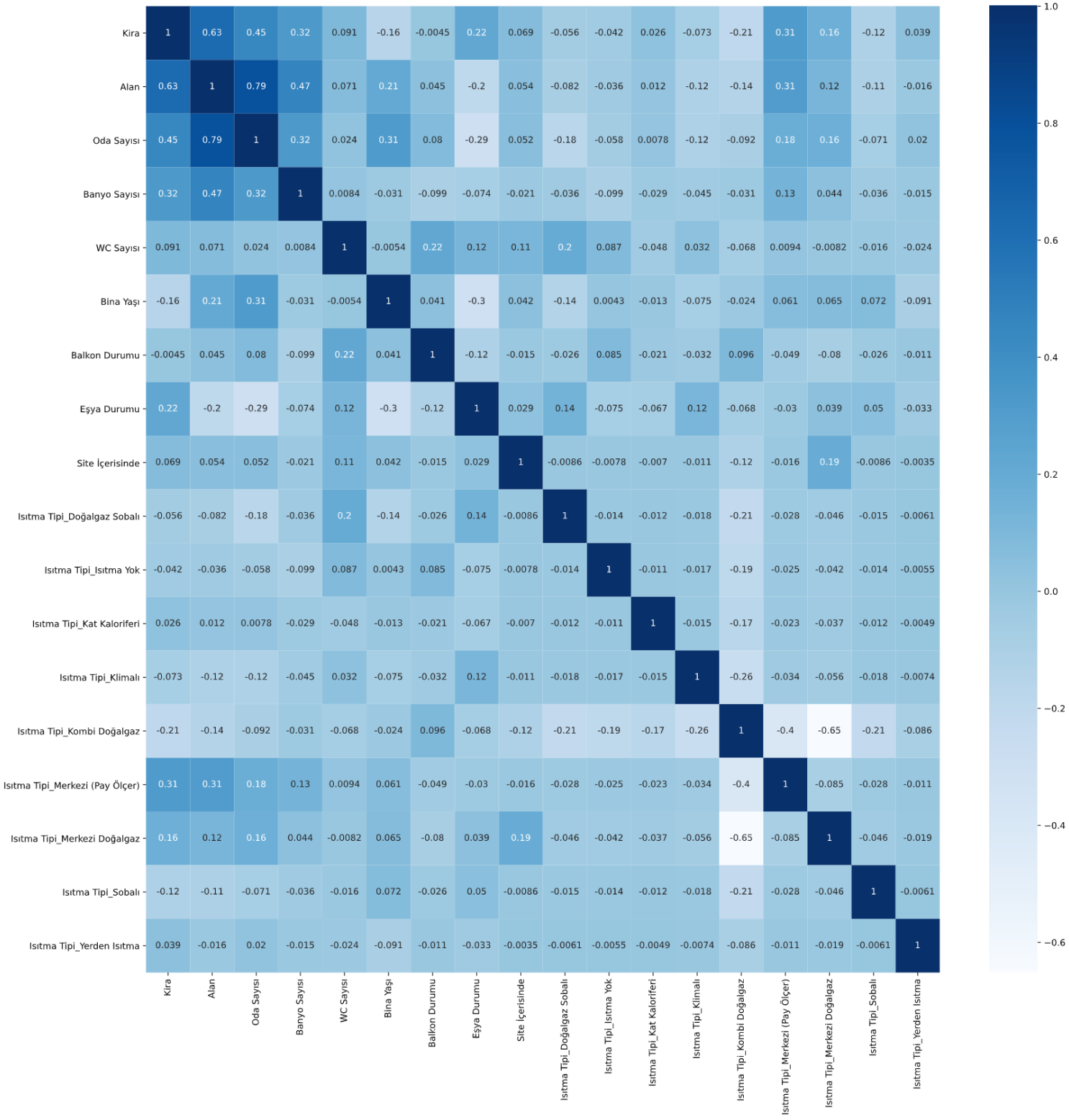
Veri kümesinde bazı abartılı ve normale uymayan örnekler bulunmaktadır. Bunun sebebi site üzerinde bazı kişilerin aylık kira kategorisine günlük ev kiralarını yazmaları ve bazı durumlarda ise satılık evlerin kiralık evler arasına karışmasıdır. Ayrıca normal kiralık ev olsa dahi çok uçuk veya çok düşük, Linear Regresyon’u olumsuz etkileyebilecek veriler de temizlenmiştir. Kirası 10000 TL üzerinde olan ve 1000 TL altında olan tüm evler veri setinden çıkarılmıştır. Veri kümesinin ilk hali ile düzenlemeler yapıldıktan sonraki son halinden bir kesit aşağıda gösterilmiştir.

	Kira	Alan	Oda Sayısı	Banyo Sayısı	WC Sayısı	Bina Yaşı	Isıtma Tipi	Balkon Durumu	Eşya Durumu	Site İçerisinde
0	4100	80	2+1	1	1	20 Ve Üzeri	Merkezi Doğalgaz	Yok	Boş	Evet
1	3100	80	3+1	1	0	20 Ve Üzeri	Kombi Doğalgaz	Yok	Boş	Hayır
2	15000	210	5+2	1	2	20 Ve Üzeri	Merkezi Doğalgaz	Var	Boş	Hayır
3	3500	110	3+1	1	1	20 Ve Üzeri	Kombi Doğalgaz	Var	Boş	Hayır
4	2900	60	Stüdyo	1	1	5-10	Doğalgaz Sobalı	Yok	Eşyalı	Hayır

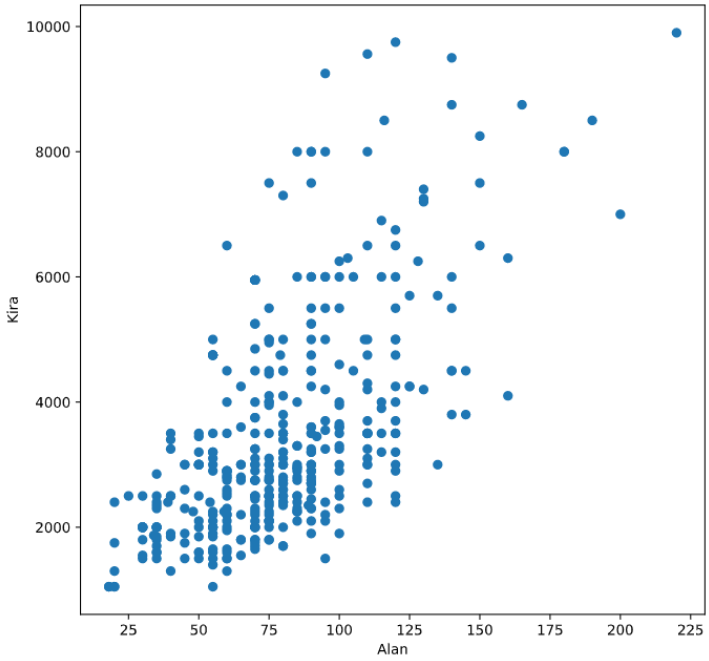
	Kira	Alan	Oda Sayısı	Banyo Sayısı	WC Sayısı	Bina Yaşı	Isıtma Tipi	Balkon Durumu	Eşya Durumu	Site İçerisinde
0	4100	80	4	1	1	25	Merkezi Doğalgaz	-1	-1	1
1	3100	80	5	1	0	25	Kombi Doğalgaz	-1	-1	-1
2	3500	110	5	1	1	25	Kombi Doğalgaz	1	-1	-1
3	2900	60	1	1	1	7	Doğalgaz Sobalı	-1	1	-1
4	2500	120	5	1	1	7	Merkezi Doğalgaz	-1	1	-1

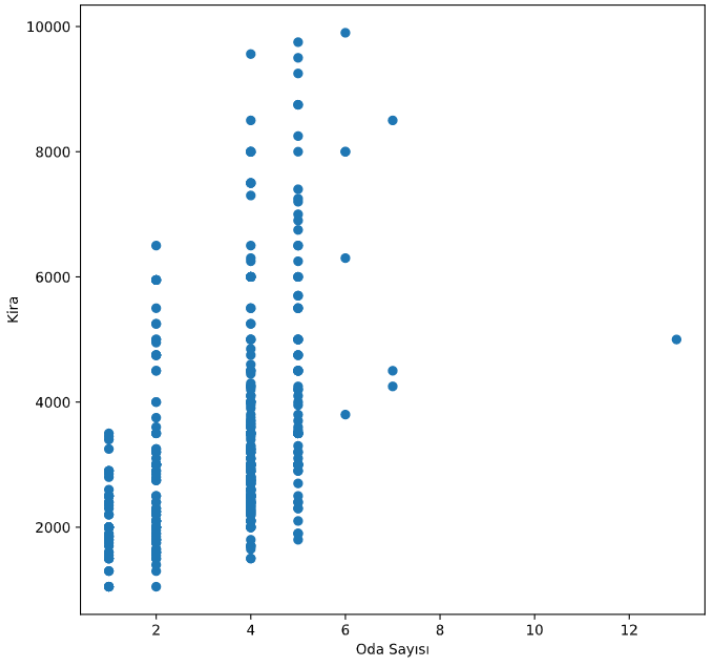
1.4 Veri Seti Görselleştirme (Data Visualization)

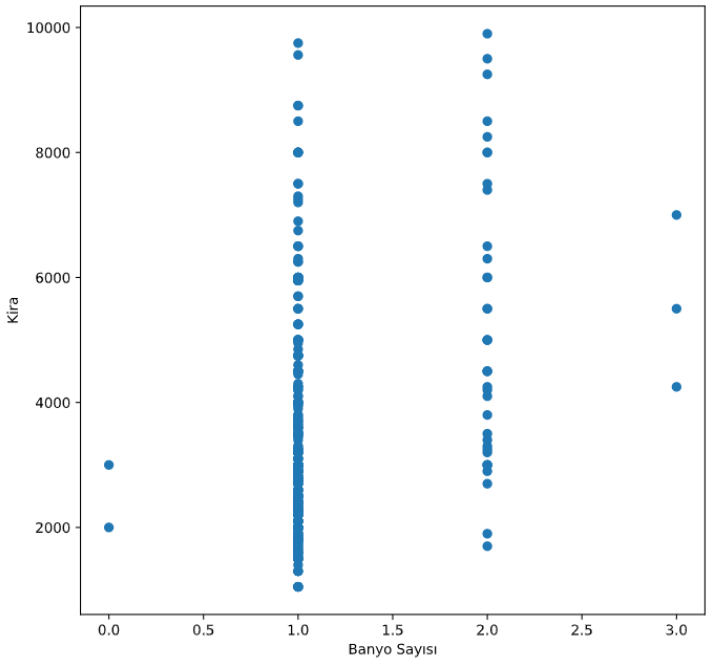
Veri kümesinin nasıl olduğunu görmek için lineer regresyon yapmadan önce çeşitli görselleştirmeler yapılmıştır. Örneğin aşağıdaki grafikte veri kümesinde bulunan özelliklerin birbirleriyle olan korelasyon ilişkisi gösterilmiştir.

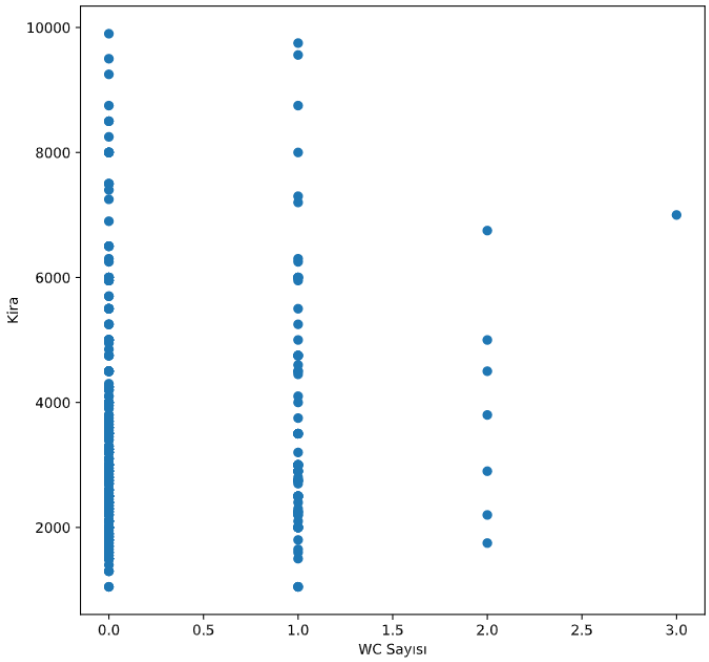


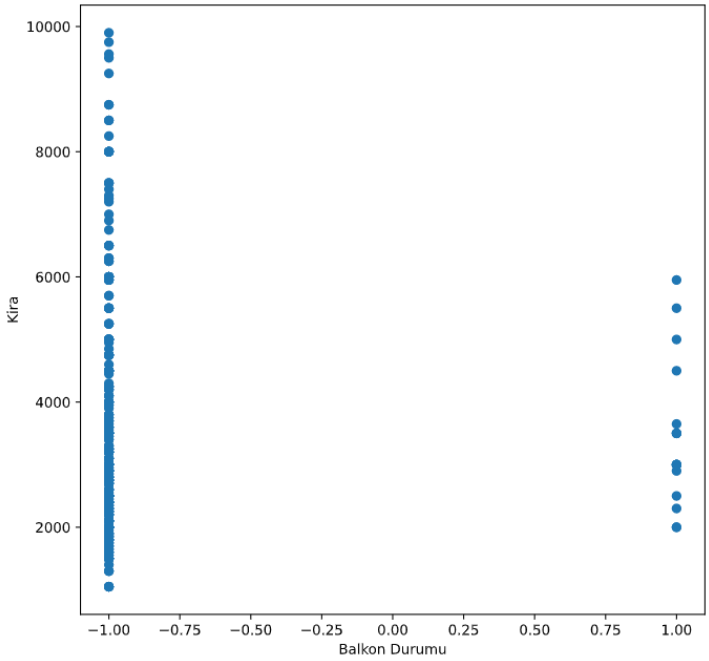
Aşağıdaki grafiklerde ise veri kümesinde bulunan bazı özelliklerin kira ile ilişkisi noktasal olarak gösterilmiştir.

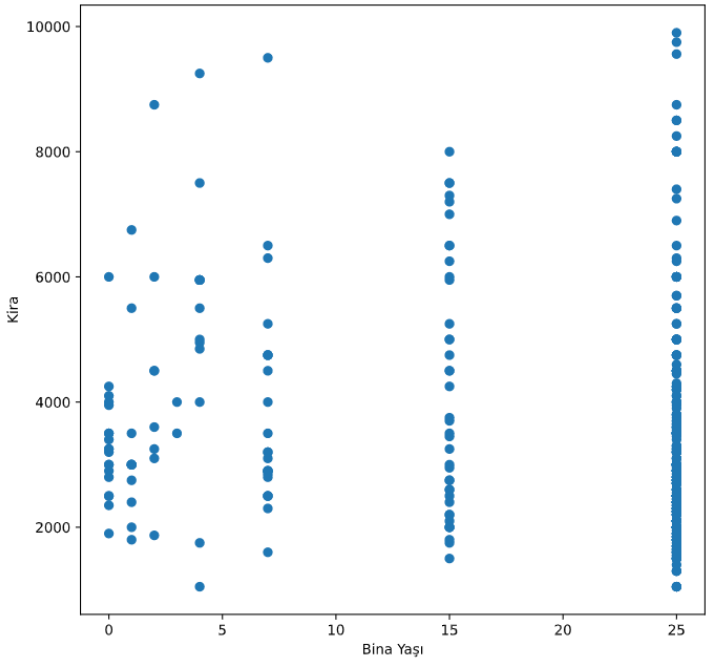


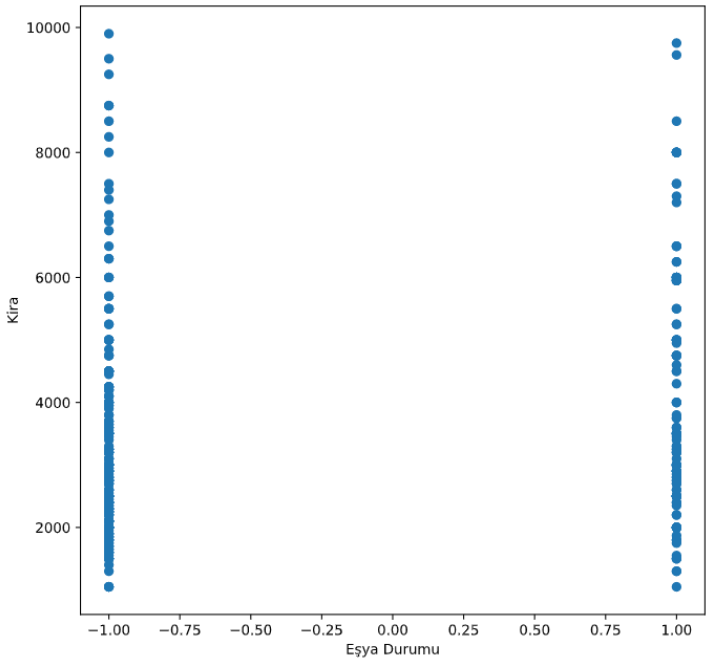


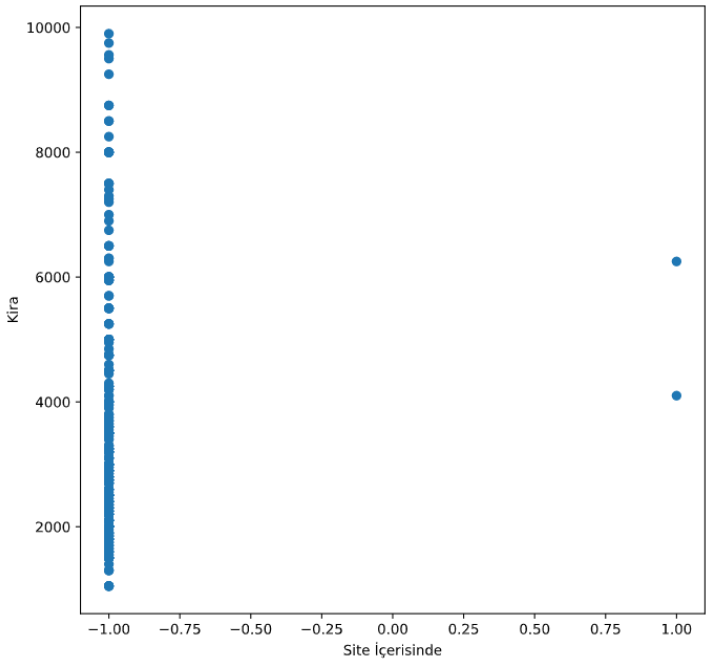





Kira
Balkon Durumu
10000
8000
6000
4000
2000
−1.00
−0.75
−0.50
−0.25
0.00
0.25
0.50
0.75
1.00


Kira
Bina Yaşı
10000
8000
6000
4000
2000
0
5
10
15
20
25

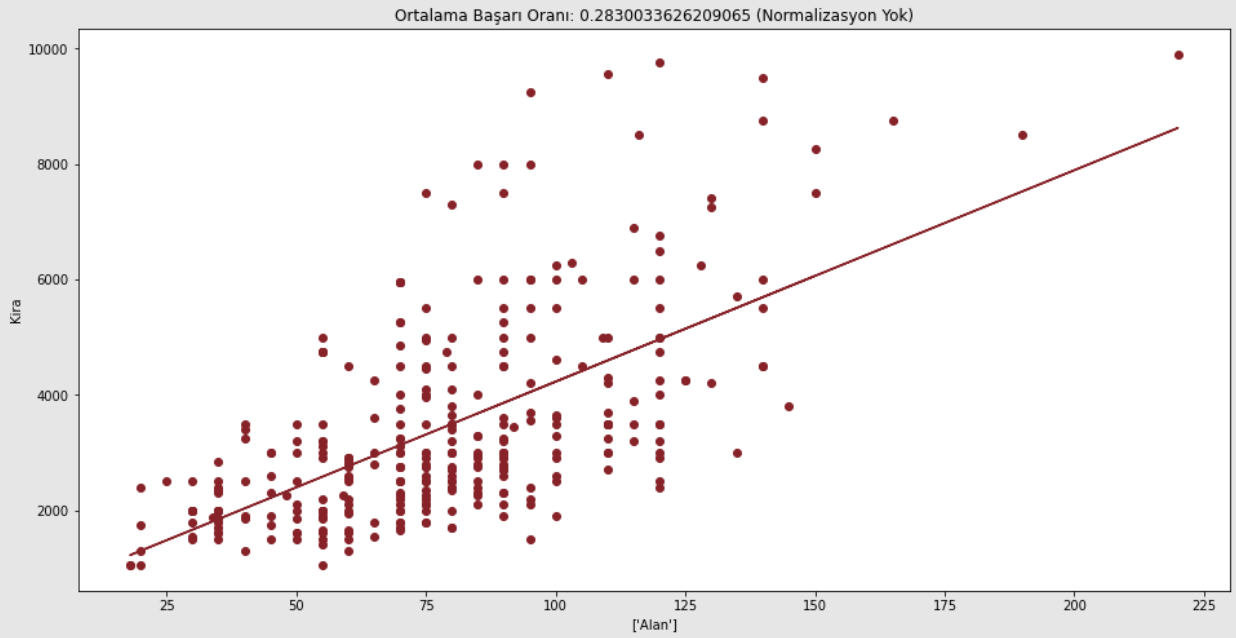

Kira
Eşya Durumu
10000
8000
6000
4000
2000
−1.00
−0.75
−0.50
−0.25
0.00
0.25
0.50
0.75
1.00

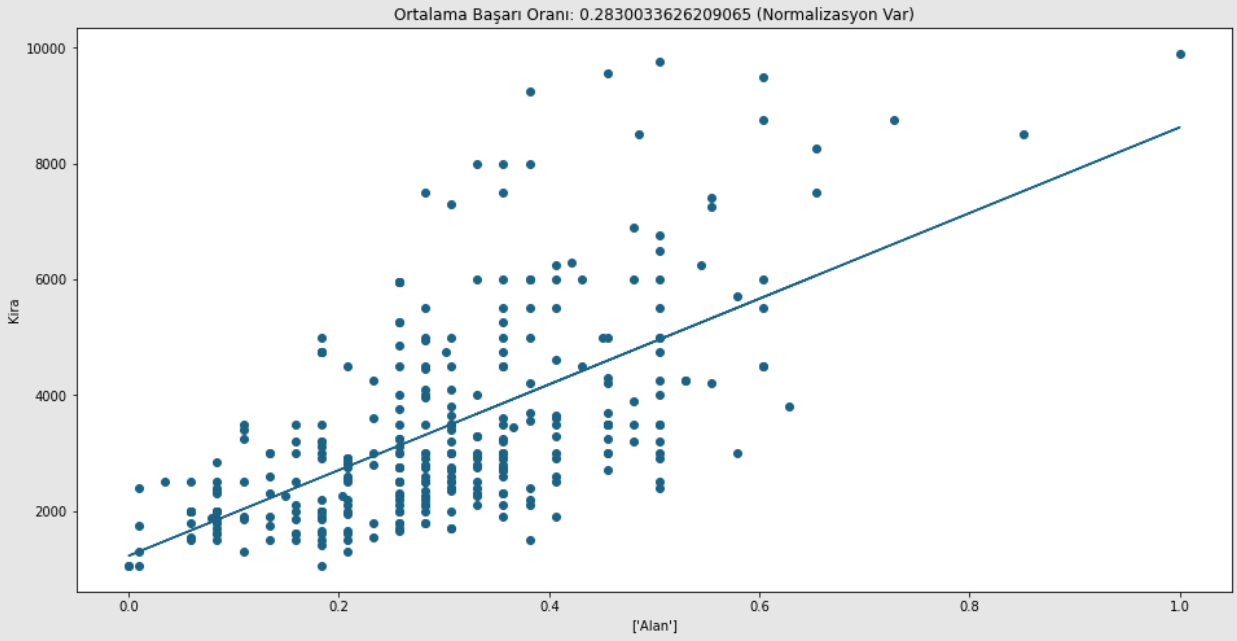

Kira
Site İçerisinde
10000
8000
6000
4000
2000
−1.00
−0.75
−0.50
−0.25
0.00
0.25
0.50
0.75
1.00

1.1 Farklı Tahminleyicilerle Deneme ve Normalizasyon

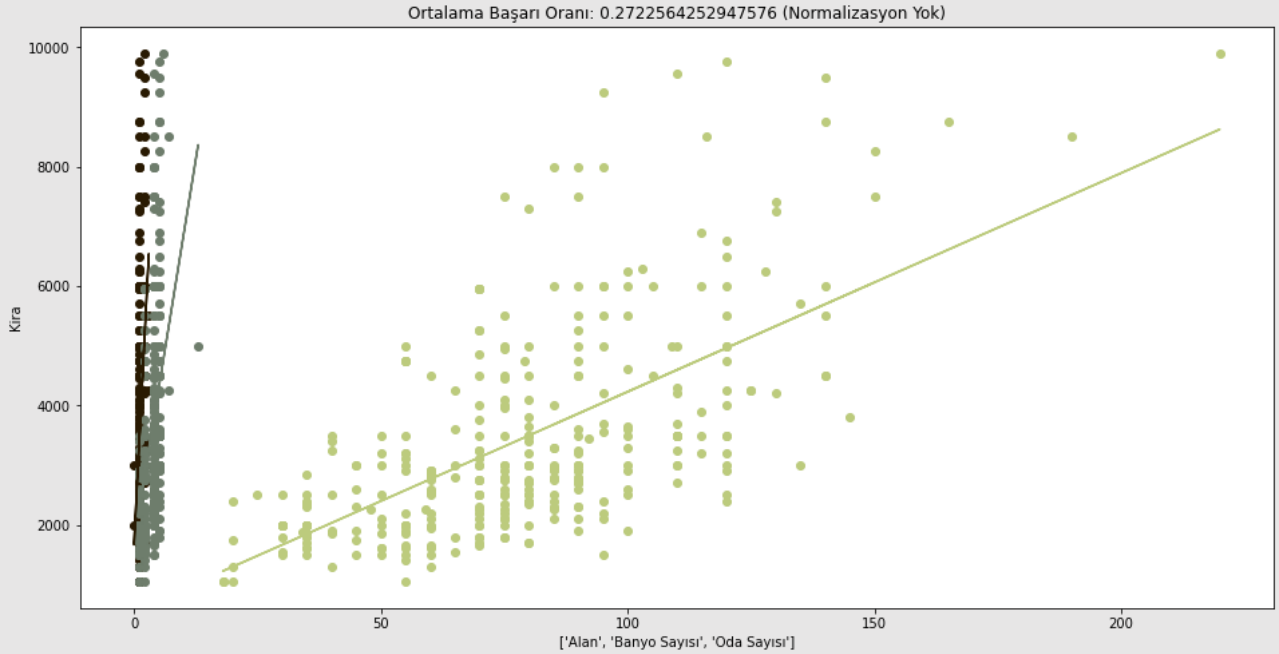
Linear Regresyon uygulamak amacıyla, 5 farklı tahminleyici oluşturulmuştur. Bu tahminleyiciler ile yapılan denemeler sonucunda 10 katlı cross validation sonuçları aşağıdaki gibidir.

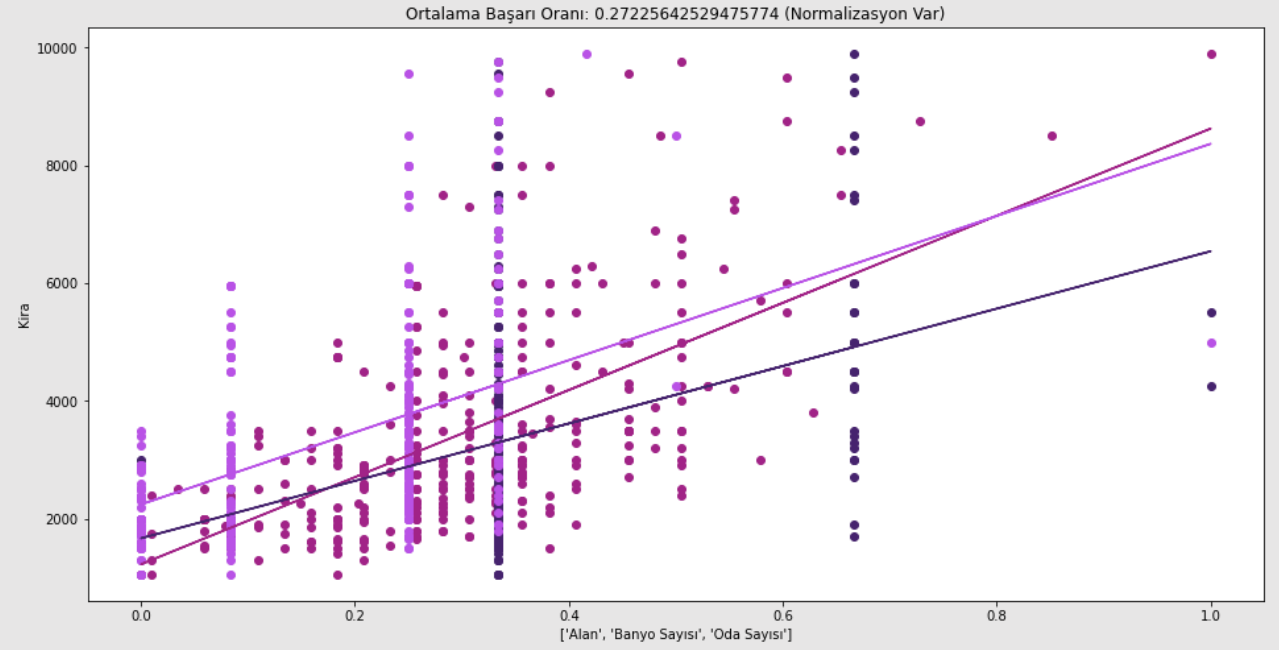
Tahminleyici 1: Kira ⇔ Alan



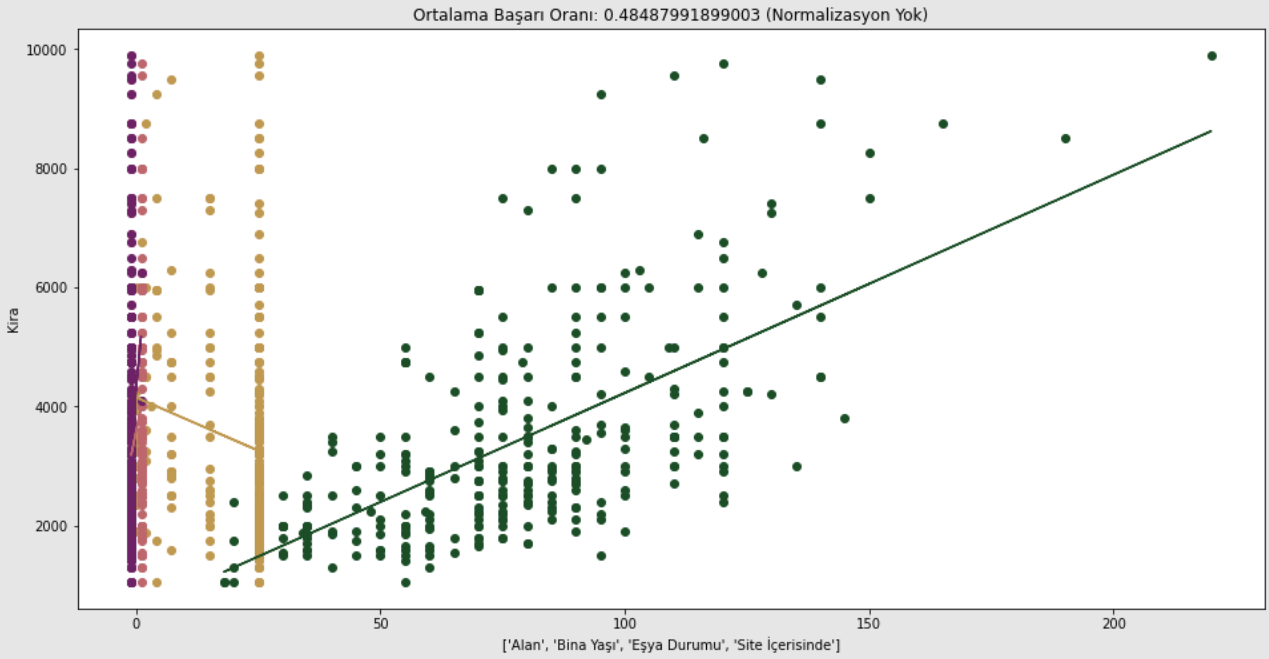


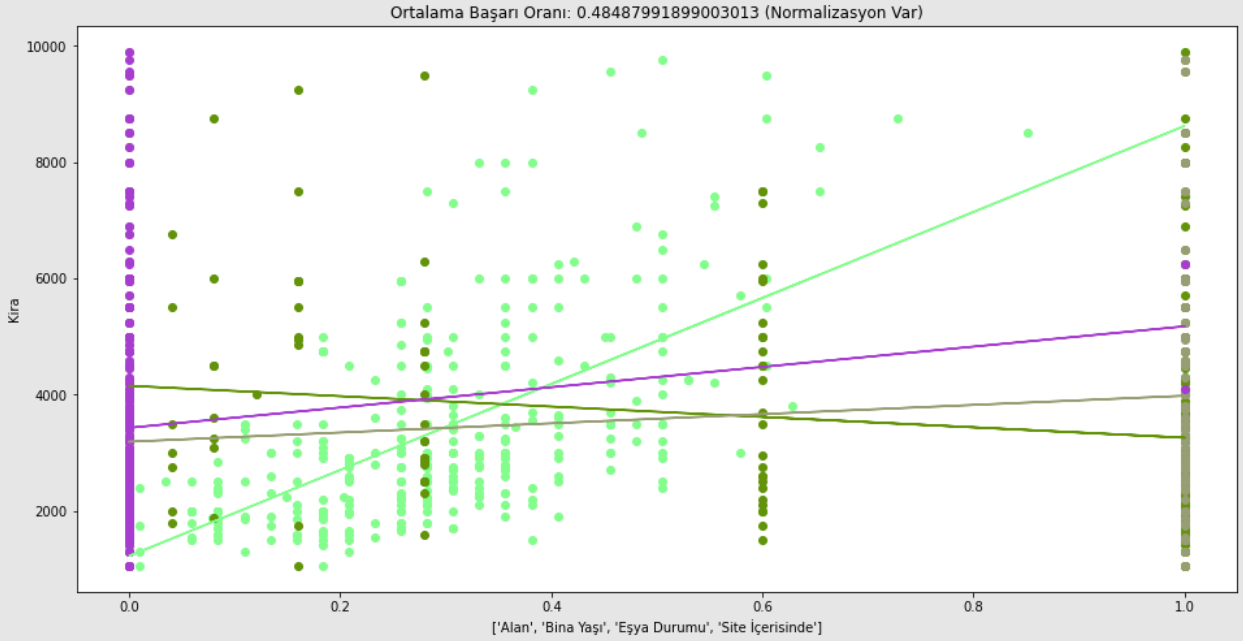
Tahminleyici 2: Kira ⇔ Alan, Banyo Sayısı, Oda Sayısı



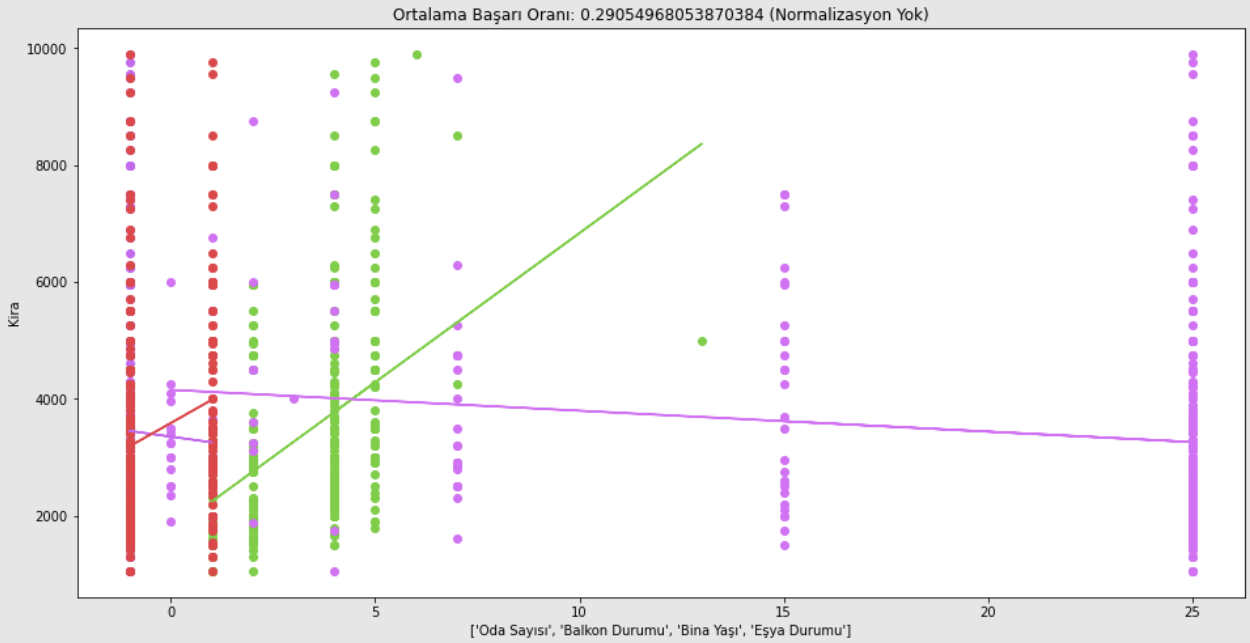


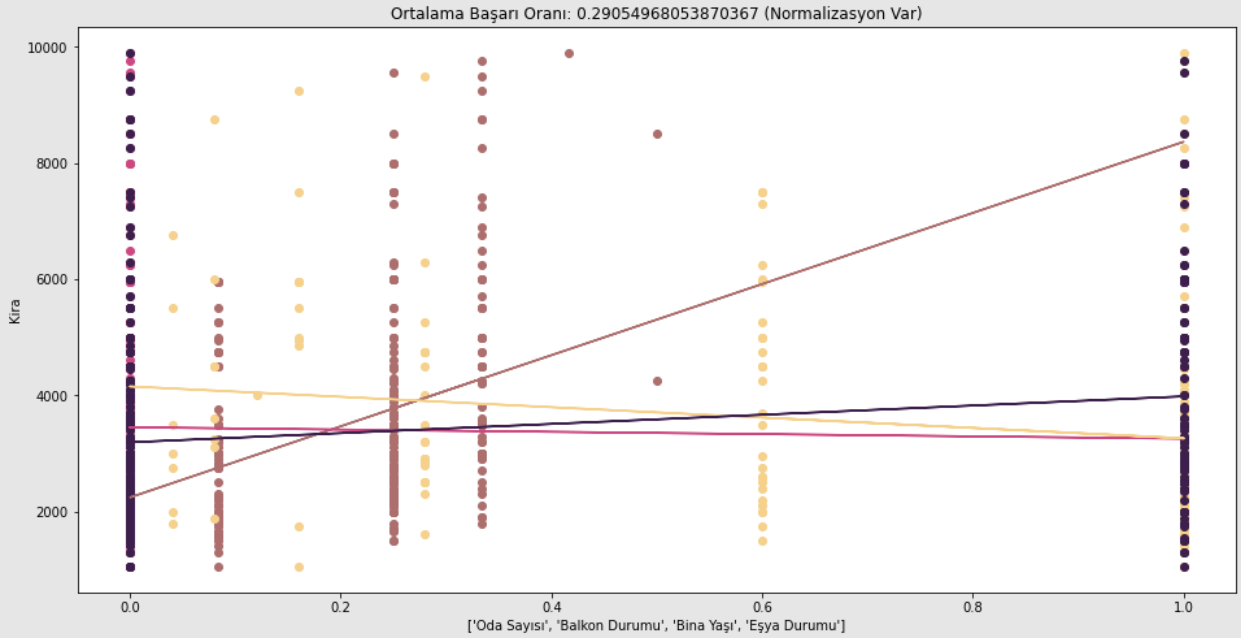
Tahminleyici 3: Kira ⇔ Alan, Bina Yaşı, Eşya Durumu, Site İçerisinde



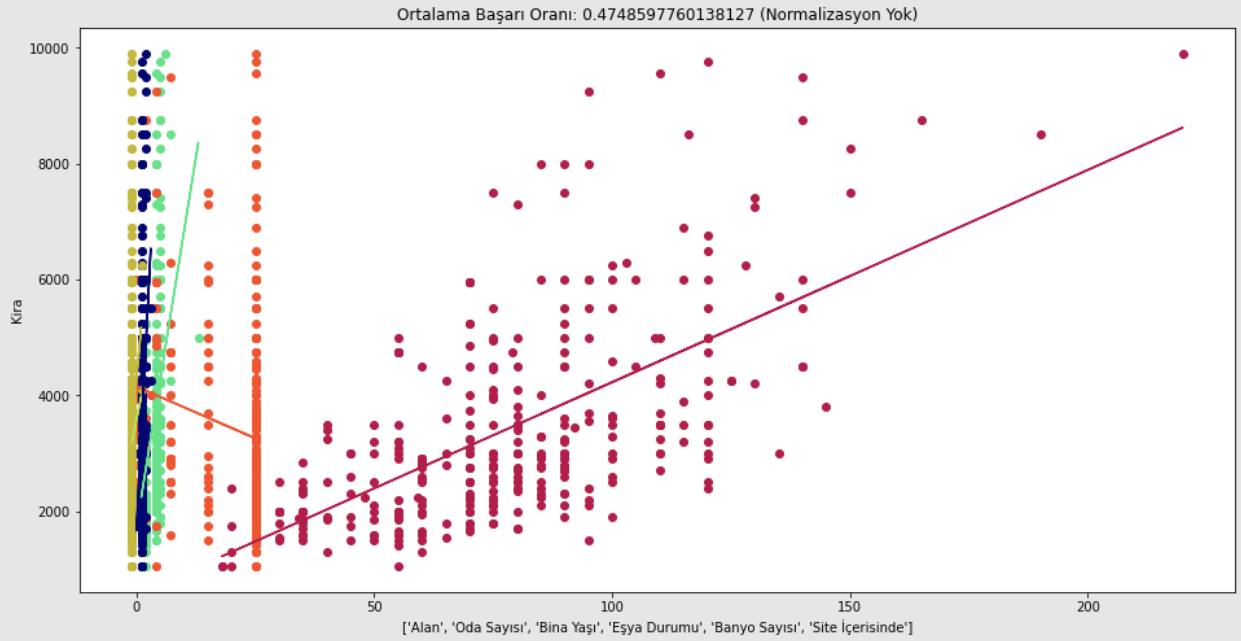


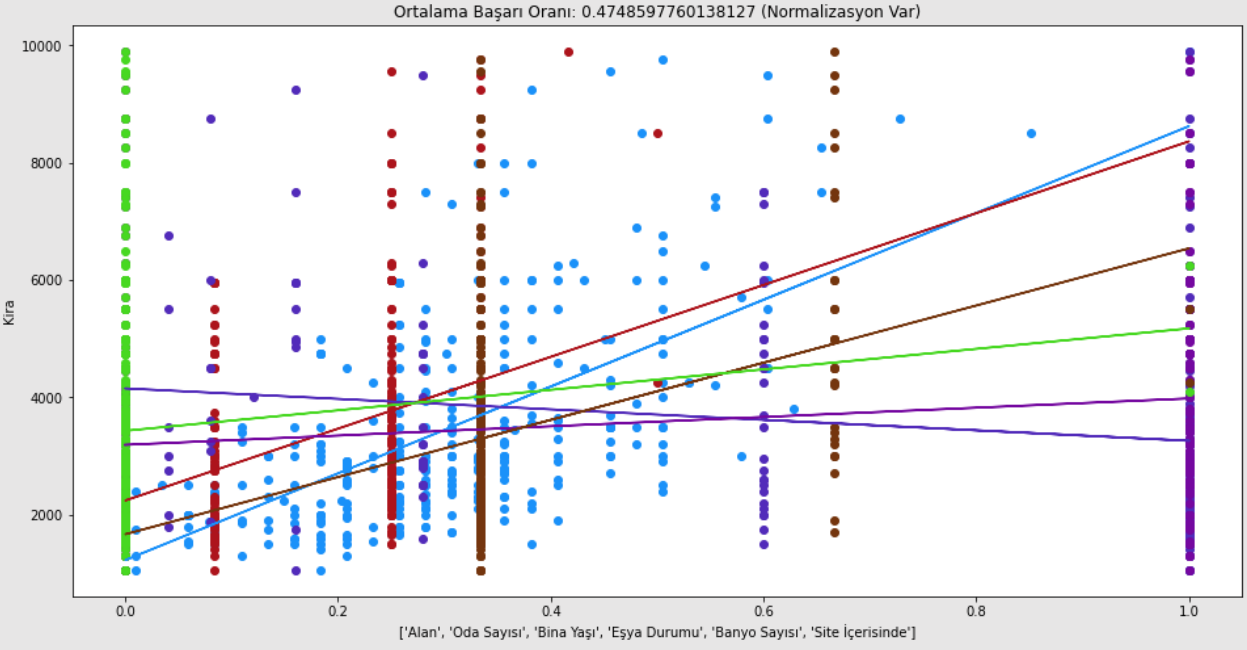
Tahminleyici 4: Kira ⇔ Oda Sayısı, Balkon Durumu, Bina Yaşı, Eşya Durumu





Tahminleyici 5: Kira ⇔ Alan, Oda Sayısı, Bina Yaşı, Eşya Durumu, Banyo Sayısı, Site İçerisinde

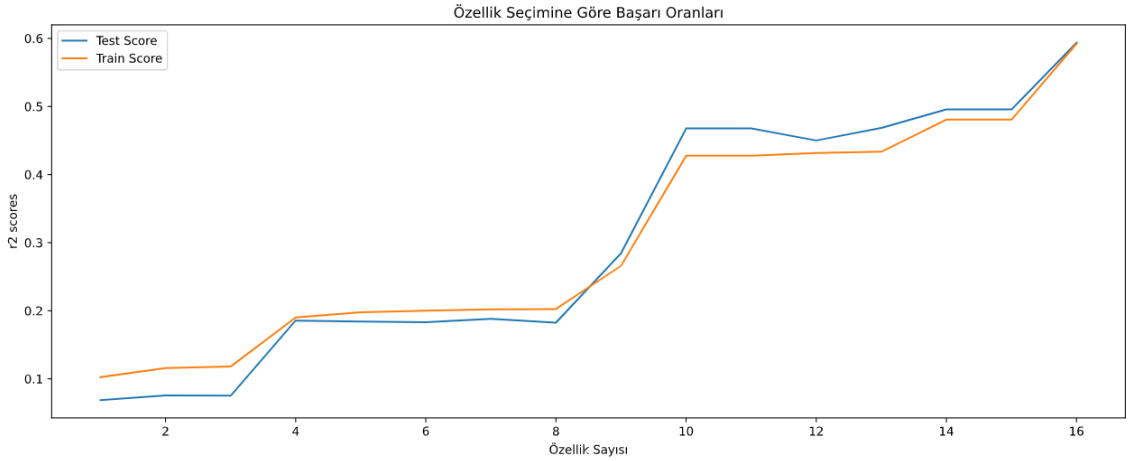




Yorum: 5 farklı tahminleyicinin normalizasyon kullanılarak ve kullanılmadanki durumlardaki sonuçları yukarıdaki gibidir. Normalizasyonun başarıya herhangi bir etkisi olmazken, ağ karmaşıklığını düşürmesiyle işlem hızlandırması ve farklı kategorilerin etkisinin beraber gözlenebilmesi adına kullanılması fayda sağlayabilmektedir. Ayrıca başarı oranı en yüksek olan tahminleyici ise 3 numaralı tahminleyici olmuştur.

1.2 Özellik Seçimi (RFE)

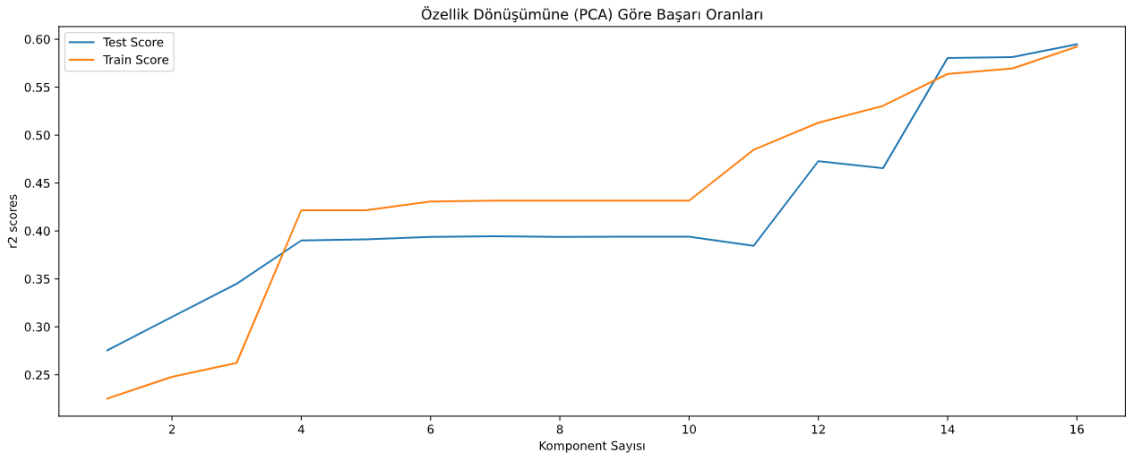
Özellik seçiminin ve özellik seçiminde kullanılan feature (özellik) sayısının etkisini daha iyi gözlemleyebilmek adına aşağıdaki gibi Özellik Sayısı – R2 Score grafiği oluşturulmuştur.



Bu grafiğe göre beklenen bir sonuç olarak özellik seçimi arttıkça başarı oranı da yükselmektedir. Özellik seçimi, başarıdan bir miktar taviz vererek, sonuca az etki eden özelliklerin elenmesi ve bu sayede sonuç hesaplamasını hızlandırması nedeniyle tercih edilebilmektedir.

1.3 Özellik Dönüşümü (PCA)

PCA özellik dönüşümünün ve seçilen komponent sayısının etkisini daha iyi görebilmek adına, Komponent Sayısı – R2 Score grafiği aşağıdaki gibi çizdirilmiştir.



Özellik dönüşümü de özellik seçiminde olduğu gibi, özellik sayısını düşürerek hesaplamayı kolaylaştırmayı amaçlamaktadır. Bu işlemde özellik seçiminde olduğu gibi bazı özellikler yok edilmez. Onun yerine verilen bir komponent sayısı parametresine göre özelliklerin etkileri daha konsantre şekilde ifade edilmeye çalışılır. Bu işlem sonucunda yeni özellikler oluşur.

Şekilde de görülebileceği gibi komponent sayısı ne kadar düşerse özelliklerin kompakt halde ifade edilmesi o kadar zorlaşacağından başarı oranı düşmektedir. Ancak pek çok uygulamada kat ve kat düşürülen özellik sayılarının başarı oranı üzerindeki küçük düşüşler kabul edilerek kullanıldığı görülmektedir. Bu durum uzun zaman alan ve özellik sayısının çok yüksek olduğu durumlarda hızı oldukça artırmaktadır.