1. Konunun Tanıtımı

Projede gerçekleştirilen konu ev fiyatlarının tahmini uygulamasıdır. Veri kümesi internetten web scraping yöntemleri ile çekilmiştir. Veri setinde bulunan özellikler fiyat, brüt metre kare, oda ve salon sayısı, kat, bina kat sayısı, net metre kare, eşyalı olup olmadığı ve bina yaşıdır. Özellikler arasındaki korelasyonlar incelenmiştir ve fiyat tahmini için regresyon yöntemi kullanılmıştır.

2. Geliştirme Sürecinde Yaşananlar

Geliştirme sürecinde ilk olarak ev ilanlarının bulunduğu siteler incelenmiştir. Bunlardan bazıları sahibinden.com, hepsiemlak.com ve emlakjet.com'dur. En güncel ve tutarlı ilanların hepsiemlak.com sitesinde olduğu düşünüldüğü için bu site tercih edilmiştir. Sonrasında web scraping yöntemleri Python teknoloji dilinin 'selenium' kütüphanesi aracılığıyla gerçekleştirilmiştir. Bunun yanında verilerin işlenmesi için 'pandas', grafik gösterimleri için 'matplotlib' ve korelasyon için 'seaborn' kütüphaneleri kullanılmıştır. Veri seti elde edildikten sonra regresyon için 'sklearn' kütüphanesinin 'LinearRegression' modeli, başarı metrikleri için 'cross_val_score', 'r2_score', 'explained_variance_score' ve 'median_absolute_error' fonksiyonları kullanılmıştır.

Veri seti internetten çekildiği için istenmeyen harfler, noktalama işaretleri ve rakamlar temizlenmelidir. String veriler sayısal değerlere dönüştürülmüştür. Boş (NaN) veriler doldurulmuştur.

Veri setinin ilk hali:

	fiyat	metre_kare	oda_salon	kat	kat_sayisi	net_m_2	esyali_mi
0	1.450.000	130	2 + 1	0	5	/ 120	Eşyalı Değil
1	1.700.000	120	2 + 1	0	15	/ 110	Eşyalı Değil
2	15.500.000	220	4 + 1	0	5	/ 200	Eşyalı Değil
3	3.850.000	165	3 + 1	1	5	/ 140	Eşyalı Değil
4	1.950.000	100	3 + 1	0	6	/ 70	Eşyalı Değil
...	...	...	...	...	...	...	...
110	3.400.000	59	1 + 1	16	23	/ 38	Eşyalı Değil
111	4.350.000	110	2 + 1	2	4	/ 90	Eşyalı Değil
112	4.200.000	45	1 + 0	0	13	/ 35	NaN
113	865.000	110	2 + 1	0	5	/ 95	Eşyalı Değil
114	890.000	85	1 + 1	0	5	/ 70	NaN

115 rows × 7 columns

Veri setinin son hali:

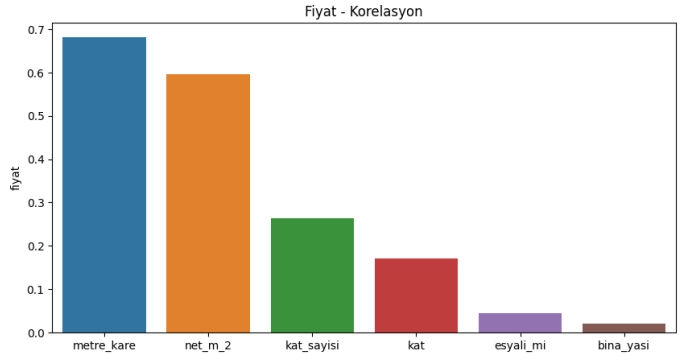
	fiyat	metre_kare	oda_salon	kat	kat_sayisi	net_m_2	esyali_mi
0	1450000	130	3	0	5	120	1
1	1700000	120	3	0	15	110	1
2	15500000	220	5	0	5	200	1
3	3850000	165	4	1	5	140	1
4	1950000	100	4	0	6	70	1
...	...	...	...	...	...	...	...
110	3400000	59	2	16	23	38	1
111	4350000	110	3	2	4	90	1
112	4200000	45	1	0	13	35	0
113	865000	110	3	0	5	95	1
114	890000	85	2	0	5	70	0

115 rows × 7 columns

3. Çeşitli Çalıştırma Örnekleri ve Sayısal Başarıları

İlk örneğimiz 71 kayıttan oluşmaktadır. (Korelasyon gösteriminde oda + salon özelliği gösterilememiştir ancak model eğitimlerinde işleme alınmıştır.)

	fiyat	metre_kare	kat	kat_sayisi	net_m_2	esyali_mi	bina_yasi
fiyat	1.000000	0.680503	0.171875	0.263817	0.595976	0.044057	0.019938
metre_kare	0.680503	1.000000	0.017863	0.166875	0.976455	0.058894	0.047412
kat	0.171875	0.017863	1.000000	0.352828	0.090797	0.062635	0.100503
kat_sayisi	0.263817	0.166875	0.352828	1.000000	0.106407	0.249613	0.017193
net_m_2	0.595976	0.976455	0.090797	0.106407	1.000000	0.079019	0.024827
esyali_mi	0.044057	0.058894	0.062635	0.249613	0.079019	1.000000	0.143063
bina_yasi	0.019938	0.047412	0.100503	0.017193	0.024827	0.143063	1.000000



Korelasyon

```
Cross-Validation Scores: [-0.07273754  1.30238563  2.71794492]
Mean Cross-Validation Score: 1.3158643363515614
Accuracy of Cross-Validation Score: % 48.41394424883652
```

Cross Validation = 3 İçin Skor

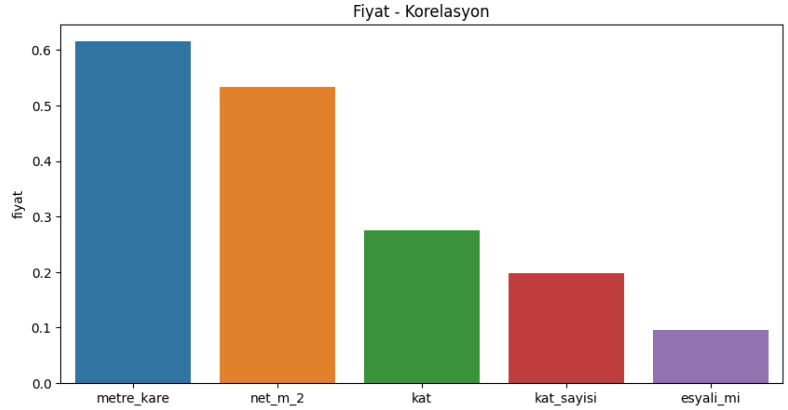
```
R-squared (R2) Score: 0.6339310354343906
Explained Variance Score: 0.6339310354343906
Median Absolute Error: 1963001.278757154
```

Başarı Oranı Metrikleri

Bu örnekte veri seti 71 kayıt içermektedir, başarı oranı yaklaşık olarak %63 olarak hesaplanmıştır.

İkinci örneğimiz 115 kayıttan oluşmaktadır. Bu veri seti bina yaşı özelliği çıkarılarak test edilmiştir.

	fiyat	metre_kare	kat	kat_sayisi	net_m_2	esyali_mi
fiyat	1.000000	0.614825	0.274793	0.198845	0.532743	0.095218
metre_kare	0.614825	1.000000	0.070744	0.026774	0.974991	0.069867
kat	0.274793	0.070744	1.000000	0.571417	0.009020	0.095833
kat_sayisi	0.198845	0.026774	0.571417	1.000000	0.052313	0.151459
net_m_2	0.532743	0.974991	0.009020	0.052313	1.000000	0.084088
esyali_mi	0.095218	0.069867	0.095833	0.151459	0.084088	1.000000



Korelasyon

```
Cross-Validation Scores: [-0.50425166  2.30939821 -0.36324909  1.88725063]
Mean Cross-Validation Score: 0.8322870193314613
Accuracy of Cross-Validation Score: % 36.03912988965356
```

Cross Validation = 4 İçin Skor

```
Cross-Validation Scores: [-0.46159271  1.48760746  1.54974228  1.92100301 -0.4771833   0.86971795]
Mean Cross-Validation Score: 0.8148824498895455
Accuracy of Cross-Validation Score: % 42.41963422894299
```

Cross Validation = 6 İçin Skor

```
R-squared (R2) Score: 0.5700370082315562
Explained Variance Score: 0.5700370082315563
Median Absolute Error: 2828304.1355007635
```

Başarı oranı yaklaşık olarak %57 olarak hesaplanmıştır.

4. Yorumlar

- Ev fiyatlarının belirlenmesinde çeşitli faktörler vardır. Bu faktörlerin ev fiyatı değerine olan ilişkisi incelenmiş olup, bağlaşımı yüksek olan bir özelliğin çıkarılmasında başarı oranının düştüğü gözlemlenmiştir.
- 2. Örnekte veri sayısı artmasına rağmen ‘oda + salon’ özelliğinin dışarıda kalmasının sisteme negatif etkisi görülmüştür.
- Veri setinin geliştirilmesi ve regresyon için önemli özelliklerin eklenmesi başarıya gözle görülür bir fayda sağlar. Örneğin bu projenin geliştirilmesi için evlerin konumu önem arz etmektedir ancak incelenmemiştir, İstanbul’dan rastgele semtlerden veriler alınmıştır.
- Sadece bir siteye bağlı kalmak yanılgıya sebep olabilir. Veri setine diğer sitelerden eşit sayıda veri çekilerek bir karma veri seti oluşturulabilir.
- Çeşitli regresyon yöntemleri kullanılarak bu problem için daha iyi başarı sonuçları sağlanabilir.

5. Yararlanılan Kaynaklar

- hepsiemlak.com
- stackoverflow.com
- scikit-learn.org/stable/modules/generated/sklearn.linear_model.LinearRegression.html
- geeksforgeeks.org/python-pandas-dataframe-corr

Mustafa Kemal Ekim

18011072