Makine Öğrenmesi Yöntemleriyle Kira Konut Bedellerinin Tahmin Edilmesi

Mehmet Ali AÇIKBAŞ
Hacettepe Üniversitesi
Enformatik Enstitüsü
m.ali.acikbas@gmail.com

Cenk GÜNGÖR
Hacettepe Üniversitesi
Enformatik Enstitüsü
cenkgngr@gmail.com

Özet – Bu projede belirli bir veri kümesini kullanarak Ankara şehrindeki konut kiralarını tahmin etmek amaçlanmıştır. Makine öğrenmesi ve veri analizi yöntemleri kullanılarak çeşitli algoritmalar geliştirilmiş ve karşılaştırmalı olarak değerlendirilmiştir. Bunların sonucunda kira bedellerinin belirli bir hata payı göz önünde bulundurularak tahmin edilebileceği gösterilmiştir. Daha iyi tahminler yapabilmek için veri zenginliğinin arttırılması ve konut kirasına etki eden diğer parametrelerin de araştırılıp hesaba katılması gerektiği sonucuna varılmıştır.

I. GİRİŞ

Hem evini kiraya vermek isteyen hem de ev kiralamak isteyen insanların yararlanabileceği aynı zamanda emlak firmalarının profesyonel olarak kullanabileceği bir kira tahmin sisteminin temellerini oluşturmak amacıyla Ankara şehrinin 10 farklı semti pilot bölgeler olarak seçilmiş ve bu semtlerdeki 300 farklı kiralık konut ilanı 'sahibinden.com' [1] platformundan faydalanılarak toplanmıştır. İlanlarda istenen kira bedeline ek olarak dairenin bulunduğu semtin adı, oda sayısı, daire büyüklüğü(metrekare cinsinden), bina yaşı, dairenin bulunduğu kat, dairenin möbleli olup olmadığı bilgileri elde edilmiştir. Ayrıca veri kümesindeki her ilan için bir ilan numarası atanmıştır. Veri kümesinden rastgele alınmış bir kesit Tablo 1'de görülebilir.

id	location	age	rooms	m2	furniture	floor	price
96	Cevizlidere	7	5	220	Evet	0	3000
177	Cukuramba	14	4	200	Evet	4	4900
197	Emek	32	4	135	Evet	0	1850
88	Cebeci	26	4	135	Evet	4	1400
214	Emek	27	4	120	Evet	1	3100
21	100. Yil	35	3	110	Evet	0	2300
274	Ovecler	21	4	100	Evet	1	1300

Tablo 1. Veri Kümesinden Bir Kesit

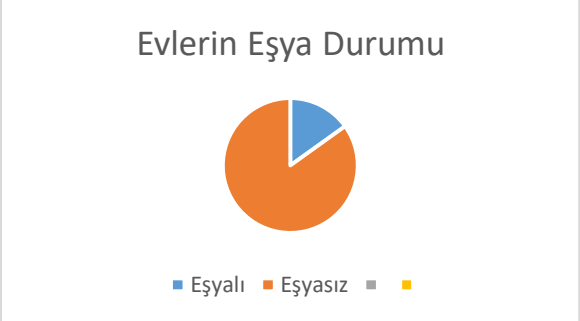
II. VERİNİN ÖN İŞLENMESİ

Elimizdeki veriden sağlıklı işleyen algoritmalar elde edebilmek için veriyi bazı ön işleme aşamalarından geçirmek gereklidir. Böylece eksik veya hatalı girilmiş veriler ayıklanabilir, veriler doğru girilmişse bile algoritmayı yanıltacak etmenlerin sayısı ve etkisi azaltılmış olur. Bu bağlamda ilk önce kira değerleri çoktan aza doğru sıralanmış ve kira bedellerinin yüzde 96'sının 750 ile 6000 TL arasında olduğu görülmüştür. Bu nedenle kira bedeli 6500 ila 15000 TL arasındaki veriler sapan veri olarak değerlendirilmiştir. Benzer şekilde daire büyüklüğü 280 metrekareden daha fazla olan daireler de sapan veri grubuna gönderilmiştir. Veri ön işleme sonucunda başta 300 olan satır sayısı 278'e düşmüştür. Son durumda bazı sayısal parametrelerin en küçük, en büyük ve ortalama değerleri Tablo 2'de görülebilir.

Parametre	Minimum	Maximum	Mean
Büyüklük(m2)	20	240	122
Kat(floor)	0	20	2.91
Kira(price)	750	600	2055
Oda sayısı(rooms)	1	5	3.406
Bina yaşı(age)	0	39	16

Tablo 2. Sayısal Parametrelerin Temel Analizi

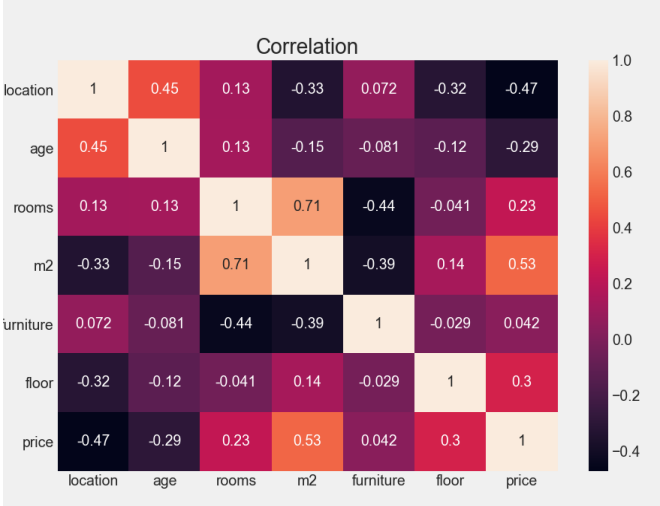
Dairelerin büyük kısmı eşyasız olarak kiraya verildiğinden ve veri setimiz güncel olduğundan möblesiz ev sayısının möbleli ev sayısına göre çok fazla olduğu Şekil 1'de görülebilir. Toplamda 42 eşyalı 236 eşyasız ilan bulunmaktadır.



Şekil 1. Evlerin Eşya Durumu

III. VERİNİN ANALİZİ

Parametrelerin ev fiyatları üzerindeki etkisini daha iyi anlayabilmek için semtler 1'den 10'a kadar numaralandırılmış ve parametreler arasında ikili korelasyon ısı grafiği çıkarılmıştır. Şekil 2'deki grafiğe bakarak dairenin bulunduğu kat, oda sayısı ve büyüklüğü arttıkça fiyatının arttığını ancak bina yaşı arttıkça kira bedelinin azaldığını söyleyebiliriz. Ayrıca dairenin bulunduğu lokasyonun kiraya çok büyük etkisi olduğu da ortadadır. Bu verileri analiz etmeden önce beklediğimiz sonuçlar da buydu. Bu grafikte bizi şaşırtan nokta ise evin eşyalı olup olmamasının fiyatlar üzerindeki etkisinin çok az olmasıydı. Bunda eşyalı ev oranının (%15.1) çok az olmasının ve eşyaların niteliğine dair elimizde bilgi olmamasının etkili olduğunu düşünüyoruz.



Şekil 2. Korelasyon Isı Grafiği

Korelasyon grafiği bize iyi bir bakış açısı kazandırsa da verileri daha ayrıntılı incelemek gereklidir. Bu nedenle ilk olarak evlerin oda sayısına göre ortalama fiyatları bulunmuştur. Tablo 3'de görüldüğü üzere tek odalı evlerin ortalama fiyatları, ortalama daire fiyatı olan 2055 liradan %36 düşükken 2 veya 3 odalı evlerin fiyatları ortalamaya çok yakındır. 4 odalı ve 5 odalı evler ise ortalamadan sırasıyla %8 ve %22 daha pahalıdır.

Oda Sayısı	Ortalama Kira
1	1325.4
2	1989.1
3	1997.8
4	2226.1
5	2527.5

Tablo 3 Odalarına Göre Evlerin Ortalama Kiraları

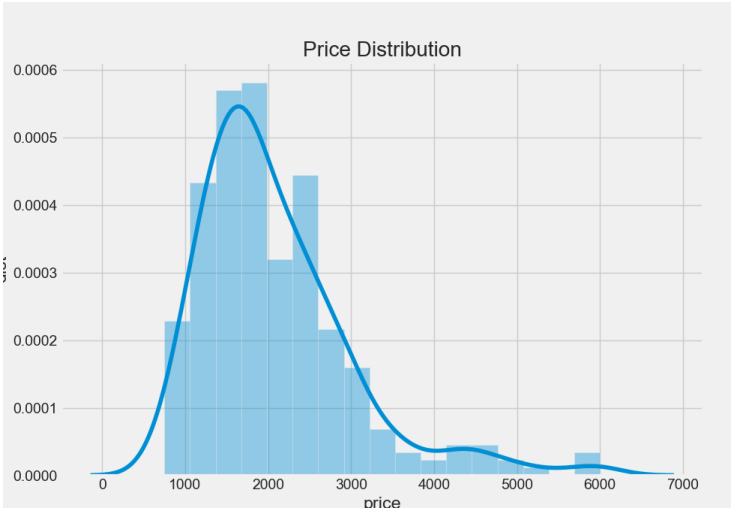
Benzer bir tablo bina yaşına göre çıkarılmıştır. Buna göre yaşı 20'nin altında olan dairelerin ortalamadan %14 daha pahalı, yaşı 20 ile 30 arasında olanların % yüzde 10 daha ucuz, yaşı 30'dan fazla olanların yüzde 30 daha ucuz olduğu görülmüştür(Tablo 4)

Bina Yaşı	Ortalama Kira
0-9	2356.5
10-19	2324.2
20-29	1844.9
30-39	1440.3

Tablo 4. Yaşlarına Göre Evlerin Ortalama Kiraları

Oda sayısı ve ev büyüklüğü arasındaki pozitif korelasyon çok yüksek olduğundan ev büyüklüğü için benzer bir analiz yapmaya gerek duyulmamıştır. Zemin katta bulunan daire fiyatlarının ortalamadan yaklaşık %11 daha ucuz olduğu görülmüştür. Eşyalı dairelerin ise ortalamadan yalnızca yüzde 2.5 daha pahalı olması sebebiyle bu parametrenin algoritmanın kararı üzerinde etkisinin neredeyse hiç olacağı aşikardır.

Şekil 3'te ise tüm dairelerin fiyat dağılımı gösterilmiştir. İlk bakışta 4000 liraya kadar olan fiyatlarda bir anormallik göze çarpmazken 4000-6000 lira arasında çok az sayıda daire olduğu gözlenebilir. Bunun sebebi özellikle Çukurambar semtinin diğer semtlerden çok daha pahalı evler içermesidir. Bu noktada tüm lokasyonların minimum, maximum ve ortalama kira değerlerinin tespit edilmesinin gerekli olduğu anlaşılmıştır.



Şekil 3. Tüm Evlerin Kira Bedeli Dağılımı

Tablo 5'te görüldüğü üzere dairelerin bulunduğu lokasyon tek başına en büyük etkiyi yaratan parametrelerden biridir. Cebeci semti fiyatların düşüklüğüyle dikkat çekerken Söğütözü ve Çukurambar ise yüksek kira bedelleriyle öne çıkmaktadır.

Semt	Minimum	Maximum	Mean
100. Yıl	1000	3000	1767.2
Bağlıca	950	4250	1943.3
Cebeci	750	2500	1370.8
Cevizlidere	1200	3000	1995.2
Çiğdem	1550	4500	2363.7
Çukurambar	2400	6000	3812.5
Emek	1150	3100	1983.3
Karapınar	950	4500	1663.4
Öveçler	900	2500	1518.3
Söğütözü	1650	3200	2455.6

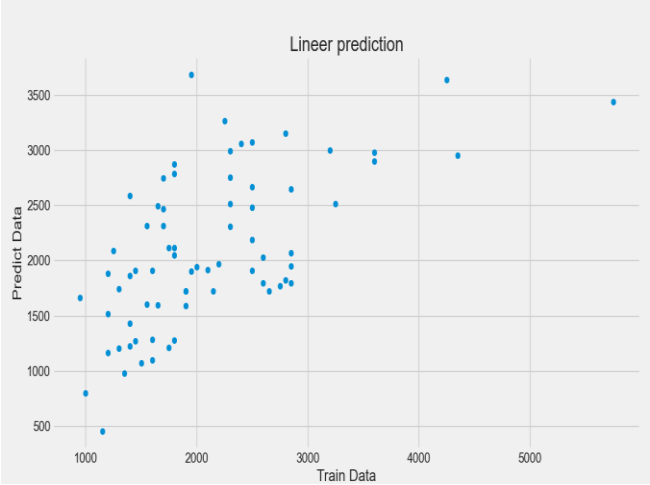
Tablo 5. Semtlerin Kira Değeri Analizi

IV. TAHMİN ALGORİTMALARI

Toplam verinin %75'i öğrenme verisi, %25'i ise test verisi olacak şekilde veri rastgele iki parçaya bölünmüştür. Veri setinin büyüklüğü ve parametrelerin etki büyüklüğünün farklı olması sebebiyle karar ağacı üzerinden kira tahmini yapma denemelerimiz modelin oldukça karmaşık olması sebebiyle başarısızlıkla sonuçlanmıştır. Bu nedenle verinin regresyon analizi Python kütüphanelerinin bazı özel fonksiyonları kullanılarak yapılmıştır. Regresyon değerlendirme metriklerinin sonuçları Tablo 6'da görülebilir. Bu metrikler MAE(Mean Absolute Error), MSE (Mean Squared Error), RMSE (Root Mean Square Error), R2 Square(Coefficient of Determination) olarak sıralanmıştır. Test verisinin lineer regresyon analiz sonuçları ise Şekil 4'teki grafikte görülebilir. Bu bilgilere bakarak lineer regresyon yönteminin kira bedellerini tahmin etmek için tek başına yeterli olmadığı görülmüştür. Bu yöntem kira bedellerini olduğundan daha fazla saptamaya eğilimlidir. Test verisinin R2_Square değerine bakarak modelin fiyatları %64 oranında yakınlıkla hesapladığını söyleyebiliriz. Bu noktada literatür araştırmaları derinleştirilmiş ve karar ağaçlarıyla regresyon modellerinin tahmini için birlikte kullanılabileceği anlaşılmıştır. [3]

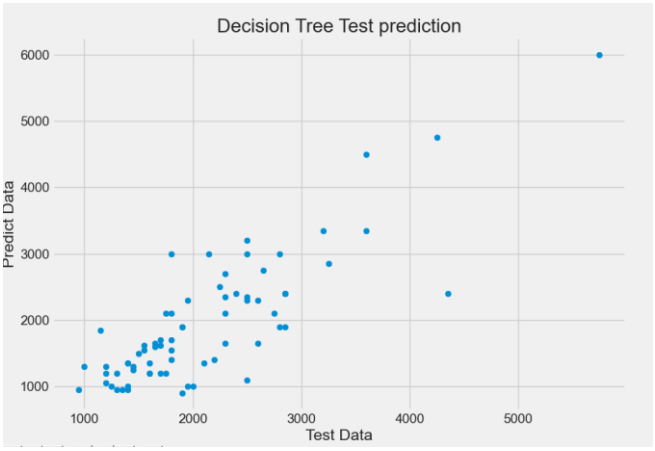
```
Test set evaluation:
_____________________________________
MAE: 370.4225352112676
MSE: 256252.81690140846
RMSE: 506.21420061216025
R2 Square 0.6379237318421661
=====================================
Train set evaluation:
_____________________________________
MAE: 0.5633802816901409
MSE: 17.370892018779344
RMSE: 4.167840210322289
R2 Square 0.9999804418614565
```

Tablo 6. Semtlerin Kira Değeri Analizi

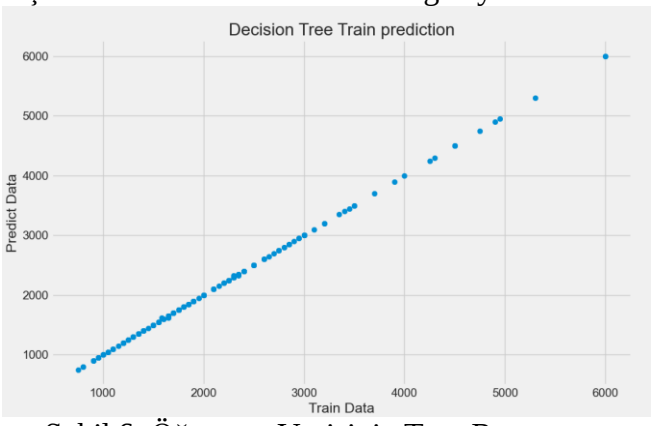


Şekil 4. Test Verisinin Lineer Regresyon Analizi

'Decision Tree Regressor' kullanılarak parametrelerin fiyat üzerindeki etki ağırlıkları daha iyi hesaba katılmış Şekil 5 ve Şekil 6'da sonuçları görülen bir model oluşturulmuştur. Bu model veri analizi bölümünde fark edilen parametrelerin fiyat üzerindeki non-lineer etkisinin karar ağacı ve regresyon modellerini birlikte kullanarak hesaplanabileceğini göstermiştir. Yüzde 10 hata payı hesaba katılarak yüzde 72 oranında, yüzde 20 hata payı hesaba katılarak yüzde 83 oranında doğru tahminler yapılabilmektedir. Test ve öğrenme kümelerinin değiştirilmesiyle bu oranlar değişebilmektedir.



Şekil 5. Test Verisinin Tree Regresyon Analizi



Şekil 6. Öğrenme Verisinin Tree Regresyon Analizi

V. SONUÇ

Yukarıdaki çalışmalar ve çıktılar sonucunda kira bedellerinin belirli verilerin işlenmesi sonucunda yazılım algoritmaları kullanılarak tespit edilebileceği görülmüştür. Sonuçlardaki hataların en önemli kaynağı olarak evlerin iç mimari özelliklerinin veri setimizde bulunmaması düşünülmektedir. Bununla beraber hata oranının azaltılması için tree regresyon modelinin çeşitli versiyonlarını araştırmak gerekir. Tüm parametreler hesaba katılsa ve daha iyi algoritmik modeller geliştirilse dahi ekonominin sosyal sebeplerinden dolayı bazı durumlarda hata payının yüksek olması kaçınılmazdır.

REFERENCES

[1] https://www.sahibinden.com/kiralik-daire/ankara-cankaya

[2] Bhaya, Wesam. (2017). Review of Data Preprocessing Techniques in Data Mining. Journal of Engineering and Applied Sciences. 12. 4102-4107. 10.3923/jeasci.2017.4102.4107.

[3] Ma, Yixuan & Zhang, Zhenji & Ihler, Alexander & Pan, Baoxiang. (2018). Estimating Warehouse Rental Price using Machine Learning Techniques. International Journal of Computers Communications & Control. 13. 235-250. 10.15837/ijccc.2018.2.3034.