KİTAP DEĞERLENDİRME PUANI VE SATIŞ MİKTARI TAHMİNİ

Recep Furkan Koçyiğit

furkan.kocyigit@std.yildiz.edu.tr

Bilgisayar Mühendisliği Bölümü

Elektrik Elektronik Fakültesi, Yıldız Teknik Üniversitesi

Özet

Bu proje kapsamında, makine öğrenmesi yöntemleri ile belirli bir kitap satış sitesi üzerindeki veriler kullanılarak müşterilerin kitaba yaptığı değerlendirme puanının ve kitabın satış miktarının tahmini hedeflenmektedir. Bu kapsamda kullanılacak veriler web kazıma yöntemi ile elde edilmiştir. İstatiksel veri analizi yöntemleri kullanılarak veri analiz edilmiş, veri ön işlemi yöntemleri kullanılarak veri eğitim için hazır hale getirilmiştir.

Değerlendirme puanı tahmini için Karar Ağacı, Naive Bayes, Rassal Orman, K En Yakın Komşuluk ve Yapay Sinir Ağları kullanılmıştır. Üzerinde çalışılan veri, kendi içerinde dağılım olarak dengesizlik gösterdiği için başarı ölçümü olarak F1 skoru tercih edilmiştir. En başarılı sonuç K En Yakın Komşuluk algoritmasından elde edilmiştir.

Kitap satış tahmininde ise Lineer Regresyon, Destek Vektörü, Karar Ağacı, Rassal Orman ve Yapay Sinir Ağları kullanılmıştır. Başarı ölçümü olarak Ortalama Hata Karesi yöntemi kullanılmıştır. En başarılı sonuç Rassal Orman algoritmasından alınmıştır.

Giriş

Akıllı cihazların artması, bulut sistemlerinin gelişmesi ve pandemi sürecinde iş yerlerinin kapalı kalmasından dolayı e-ticarette artış meydana gelmiştir. Bir ürünün müşteriler tarafından satışının artmasında başka müşterilerin yaptığı değerlendirmelerin büyük önem arz ettiği bilinmektedir. Bu çalışmada kitap satışlarında yapılan değerlendirmelerin etkisi araştırılmış ve yapılan değerlendirmenin kitaba bağlı özelliklere bağlı olup olmadığı tespit edilmek istenmiştir.

Bu kapsamda yapılan çalışmalarda problem kullanıcı benzerliği, duygu özellikleri ve kitabın özelliklerini kullanarak oluşturulan performans skorunu bulan bir regresyon problemi olarak ele aldığı görülmektedir [1][2].

"Yelp" isimli şirketin eğitimsel amaçlarla kullanılmak üzere paylaştığı işletmeler, kullanıcılar ve kullanıcı yorumları içeren veri kümesinde yapılan çalışmalarda benzer kullanıcıların aynı değerlendirmeleri yapmasının daha muhtemel olabileceğinden dolayı kullanıcı benzerliği, ürünlere dair özellikler ve kullanıcının yaptığı yorumdan duygu analizi yapılarak bu skorunda verilen değerlendirme skoruna etkisi araştırılmıştır [1].

Bir diğer araştırmada kitap arşivi ve öneri sitesi olan "goodreads.com" üzerinde kitaplara ait bilgiler ve kullanıcıların yaptığı değerlendirmeler üzerinden bir tahmin söz konusudur. Tahmin algoritması olarak yapay sinir ağları kullanılmıştır. Kitabın ortalama değerlendirme skorunu bulmak üzerine yapılmış bir çalışmadır [2]. Yapay sinir ağı modeli, 0,005 hata ile kitapların genel oranını %99,78 doğrulukla tahmin edebilmiştir.

Kitap satış miktarı için yapılan çalışmada ise "Amazon.com" üzerinde satılan kitap ve işletme bilgileri ve kullanıcı değerlendirmeleri kullanılarak regresyon modelleri, karar ağaçları ve yapay sinir ağları ile bir tahminleme yapılmıştır [3].

Sistem Tasarımı

Proje kapsamında kullanılan veri "kitapyurdu.com" sitesinde en çok satılan 947 edebiyat ana kategorisindeki kitaplara aittir. Öncellikle en çok satılan 947 kitabın linki elde edilmiştir. Daha sonra web kazıma yöntemiyle

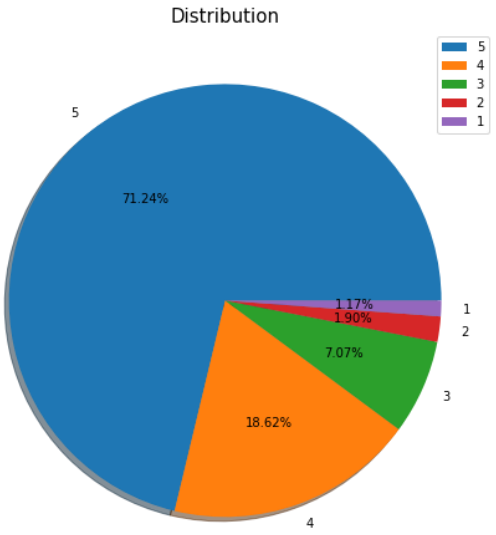
bu kitaplara ait bilgiler elde edilmiştir. Web kazıma işlemi sonucunda 35964 adet müşteri değerlendirilmesi elde edilmiştir.

Kullanılan veri kümesinin özellikleri aşağıdaki gibidir:

1. **id**: Kitabın id bilgisi
2. **book**: Kitabın ismi
3. **writer**: Kitabın yazarı
4. **publisher**: Kitabı yayınlayan yayın evi
5. **num_pages**: Kitabın sayfa sayısı
6. **num_comments**: Kitaba yapılan yorum sayısı
7. **num_favorites**: Kitabı favorilerine ekleyen kullanıcı sayısı
8. **num_will_read**: Kitabı "Okuyacağım" olarak işaretleyen kullanıcı sayısı
9. **num_is_read**: Kitabı "Okuyorum" olarak işaretleyen kullanıcı sayısı
10. **num_did_read**: Kitabı "Okudum" olarak işaretleyen kullanıcı sayısı
11. **num_purchases**: Kitabı satın alan kişi sayısı
12. **num_ratings**: Kitaba herhangi bir değerlendirme veren kişi sayısı
13. **avg_rating**: Kitabın aldığı ortalama değerlendirme puanı
14. **num_five_star**: Kitaba değerlendirme olarak 5 puan veren kullanıcı sayısı
15. **num_four_star**: Kitaba değerlendirme olarak 4 puan veren kullanıcı sayısı
16. **num_three_star**: Kitaba değerlendirme olarak 3 puan veren kullanıcı sayısı
17. **num_two_star**: Kitaba değerlendirme olarak 2 puan veren kullanıcı sayısı
18. **num_one_star**: Kitaba değerlendirme olarak 1 puan veren kullanıcı sayısı
19. **comment**: Kitaba kullanıcı tarafından yapılan yorum
20. **price** : Kitabın ücreti
21. **type**: Kitabın türü
22. **positivity**: Kitaba yapılan yorumun pozitiflik yüzdesi
23. **star**: Kitaba kullanıcı tarafından verilen değerlendirme puanı

Elde edilen ham veri üzerinde bazı sayısal özelliklerin gösteriminde farklılıklar olduğu için o alanlar düzeltilmiş, aynı isimde ama ciltli veya ciltsiz olarak kitap isminin yanında var olan bilgiler temizlenmiş, yorumlarda alfa-nümerik olmayan karakterler temizlenmiş, kitabı "Okuyorum", "Okuyacağım" ve "Okudum" olarak işaretleyen kişilerin ortalama bilgisi elde edilmiş, kitaba verilen değerlendirmelerin ortalaması alınmış, kitabın türü kategorik bir veri olduğundan One-Hot Encoding işlemi uygulanmış ve eğitilmiş bir model kullanılarak yorumlardan duygu çıkarımı yapılmıştır.

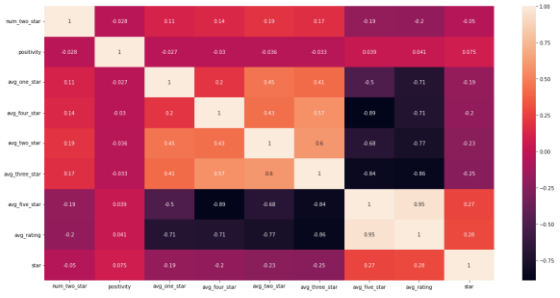
Veri içerisindeki verilen değerlendirme puanın dağılımı aşağıdaki gibidir:



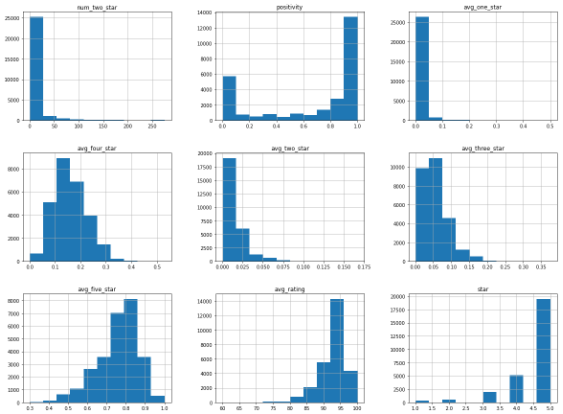
Verinin Dağılımı

Veri üstteki grafikte görülebileceği gibi eşit olarak dağılmamıştır. En çok satılan kitaplardaki müşteri değerlendirmeleri olduğundan değerlendirmelerin genel olarak 5 üzerinden 5 aldığı görülmüştür.

Değerlendirmenin tahmin edilmesi için değerlendirme 0.05 ten yüksek korelasyona sahip özellikler kullanılmıştır. Korelasyon matrisi aşağıdaki gibidir:

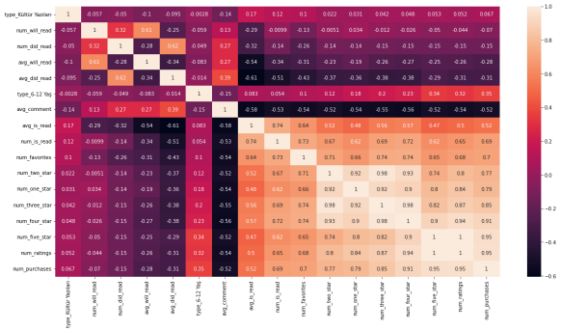


Değerlendirme puanı tahmini için kullanılan özelliklerin kendi içerisinde dağılımı ise aşağıdaki gibidir:

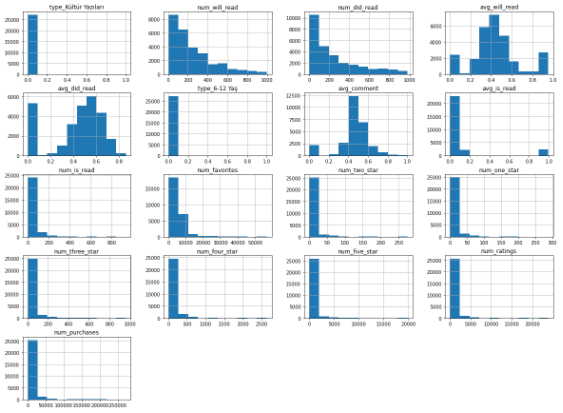


Değerlendirme puanı ile korelasyonu 0.05'ten yüksek 6 özellik kalmıştır. Kitabın fiyatının, türünün veya kitabın satış sayısının değerlendirme skoru ile ilişkisinin zayıf olduğu görülmüştür.

Kitap satış tahmini için de korelasyonu 0.05'ten büyük özellikler seçilmiştir. Seçilen özelliklerin satış miktarı ile ilişkisini gösteren korelasyon matrisi aşağıdaki gibidir:



Bu özelliklerin kendi içerisindeki dağılımı ise aşağıdaki gibidir:



Satış tahmini için bazı kitap türlerinin satış miktarı ile ilişkisinin kuvvetli olduğu gözlemlenmiştir. Ayrıca, kitap satış miktarının verinin %75'i için 10.000'nin altında kaldığı gözlemlenmiştir.

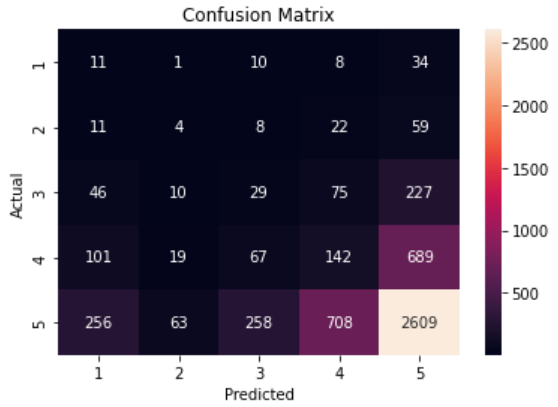
Sınıflandırma Modelleri

1. **Naive Bayes:** Bayes Teoremini kullanarak sınıflandırma yapar. Olasılıksal bir yaklaşım kullandığından verinin normalize edilmesine gerek yoktur.
2. **K En Yakın Komşuluk:** Tahminleme yaparken uzaklık kullandığından verinin normalize edilmesi gereklidir. Veri kümesine yeni eleman geldiğinde en yakın grubu bulmaya çalışır. Eğitim süreci yoktur. K değeri, çift bir sayı seçilmesinde test edilecek değerin gruplara eşit uzaklıkta çıkabilmesinden dolayı tek sayı seçilmelidir.
3. **Karar Ağacı:** Bilgi kazanımının en yüksek olduğu özelliği bulma üzerine kurulu bir algoritmadır. Özellik çıkarımına veya normalizasyona gerek yoktur. Önemli özellikler zaten seçilmiş olacaktır.
4. **Yapay Sinir Ağları:** İnsan beyninin çalışma prensibinden esinlenerek geliştirilmiştir. Normalizasyon işlemi gereklidir. Girdi katmanı, gizli katman ve çıktı katmanından oluşur. Girdi katmanından gelen değerler ile ağırlıklarının nokta çarpımının bir aktivasyon fonksiyonundan çıktısı sonucunda bir sonraki katmandaki girdi elde edilmiş olur.
5. **Rassal Orman:** Eğitim aşamasında çok sayıda karar ağacı oluşturarak tahmin yapan bir toplu öğrenme yöntemidir.

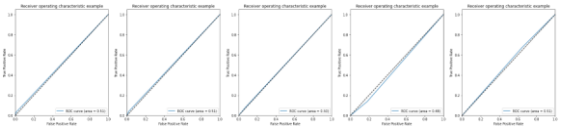
Deneysel Analiz

Karar ağacı algoritması criterion: [gini, entropy], max_depth: [5, 6, 7, 8, 9, 10, 11, 12, None],max_features: [auto, sqrt, log2] parametreleri arasında "Stratified 10 Fold Cross Validation" ile sınandığında en iyi sonucu "criterion: gini, max_depth: None, max_features: auto" parametresinden almaktadır. Bu parametreler ile eğitilen

modelin karmaşıklık matrisi ve auc grafikleri aşağıda verilmiştir.

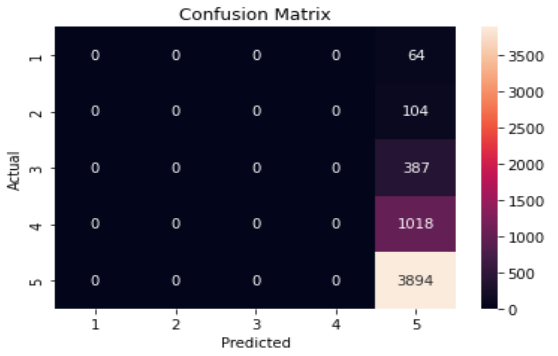


	precision	recall	f1-score	support
0	0.04	0.03	0.04	64
1	0.04	0.04	0.04	104
2	0.08	0.07	0.08	387
3	0.15	0.14	0.14	1018
4	0.72	0.67	0.69	3894
micro avg	0.55	0.51	0.53	5467
macro avg	0.21	0.19	0.20	5467
weighted avg	0.55	0.51	0.53	5467
samples avg	0.51	0.51	0.51	5467



Kullanılan veri dengesiz olduğundan başarı metriği olarak f1-skor kullanılmıştır. Sonuçlara bakıldığında algoritmanın 1, 2, 3 ve 4 değerlendirme skorları için öğrenemediği gözlemlenmektedir.

Naive Bayes algoritmasında en uygun hiper parametreleri seçmek için öncelikle model eğitim kümesiyle eğitilmiş daha sonra "alpha" parametresi [0, 0.1, 0.2, …, 1] arasındaki değerler için validasyon verisiyle "Stratified 10 Fold Cross Validation" ile sınandığında en başarılı sonucu veren "alpha" değerinin 0.0 olduğu görülmüştür.



	precision	recall	f1-score	support
0	0.00	0.00	0.00	64
1	0.00	0.00	0.00	104
2	0.00	0.00	0.00	387
3	0.00	0.00	0.00	1018
4	0.71	1.00	0.83	3894
micro avg	0.71	0.71	0.71	5467
macro avg	0.14	0.20	0.17	5467
weighted avg	0.51	0.71	0.59	5467
samples avg	0.71	0.71	0.71	5467

Naive Bayes algoritmasında 1, 2, 3 ve 4 için hiçbir öğrenme görülmemiştir. Gelen tüm girdiye karşı 5 çıktısını vermektedir.

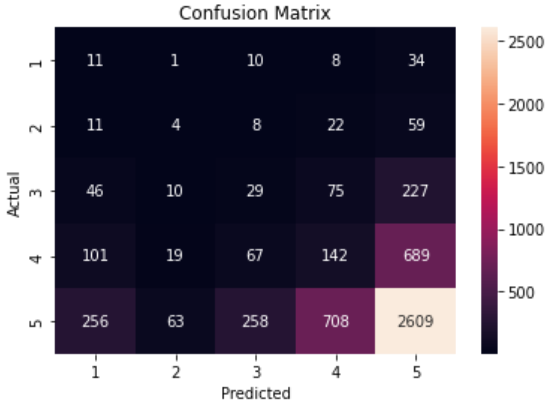
Rassal Orman Algoritması için aşağıdaki parametreler arasından en iyi model seçimi yapılmaya çalışılmıştır:

bootstrap: [True, False], max_depth: [5, 6, 7, 8, 9, None], max_features: [auto, sqrt], n_estimators: [100, 200, 300]

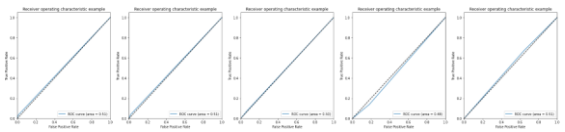
En başarılı sonucu veren parametreler ise şu şekildedir:

bootstrap: False, max_depth:None, max_features:auto, n_estimators: 100

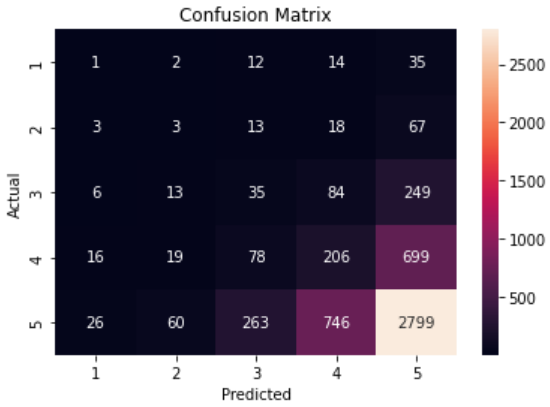
Elde edilen sonuçlar ise aşağıdaki gibidir:



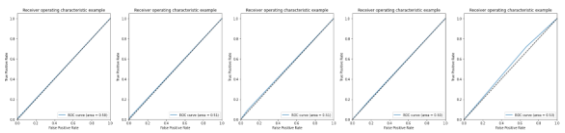
	precision	recall	f1-score	support
0	0.04	0.03	0.04	64
1	0.04	0.04	0.04	104
2	0.08	0.07	0.08	387
3	0.15	0.14	0.14	1018
4	0.72	0.67	0.69	3894
micro avg	0.55	0.51	0.53	5467
macro avg	0.21	0.19	0.20	5467
weighted avg	0.55	0.51	0.53	5467
samples avg	0.51	0.51	0.51	5467



K en yakın komşuluk algoritmasında komşu sayısı için 1 ile 15 arasındaki komşu değerleri için validasyon verisi ile test edilmiş ve en iyi sonuç komşu sayısı 1 iken elde edilmiştir.



	precision	recall	f1-score	support
0	0.02	0.02	0.02	64
1	0.03	0.03	0.03	104
2	0.09	0.09	0.09	387
3	0.19	0.20	0.20	1018
4	0.73	0.72	0.72	3894
micro avg	0.56	0.56	0.56	5467
macro avg	0.21	0.21	0.21	5467
weighted avg	0.56	0.56	0.56	5467
samples avg	0.56	0.56	0.56	5467



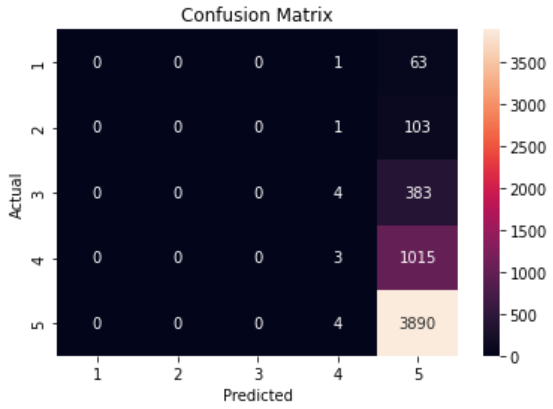
Grafiklerden görülebileceği üzere f1 skoru diğer 2 algoritmadan daha yüksektir.

Yapay sinir ağlarında ise model şu şekilde oluşturulmuştur:

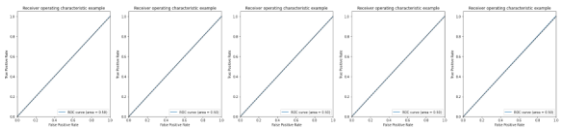
```
def build_model():
    model = Sequential()
    model.add(Dense(80, kernel_initializer = 'uniform', activation = 'relu', input_dim = X_train_standardized.shape[1]))
    model.add(Dense(40, kernel_initializer = 'uniform', activation='relu'))
    model.add(Dense(20, kernel_initializer = 'uniform', activation='sigmoid'))
    model.add(Dense(5, activation="softmax"))
    model.compile(loss="categorical_crossentropy", optimizer="adam", metrics=["accuracy"])

    return model
```

Modelden elde edilen sonuçlar ise şu şekildedir:



	precision	recall	f1-score	support
0	0.00	0.00	0.00	64
1	0.00	0.00	0.00	104
2	0.00	0.00	0.00	387
3	0.23	0.00	0.01	1018
4	0.71	1.00	0.83	3894
micro avg	0.71	0.71	0.71	5467
macro avg	0.19	0.20	0.17	5467
weighted avg	0.55	0.71	0.59	5467
samples avg	0.71	0.71	0.71	5467



Tüm modellere bakıldığında başarılı bir tahmin yapabilecek model görülmemektedir. AUC çok nadir olarak ve genelde 5. Değerlendirme için 0.5'nin üzerine çıkmıştır. Bunun sebebinin kitaba ait özelliklerin müşterinin değerlendirme puanı vermesinde etkisinin düşük olmasıdır.

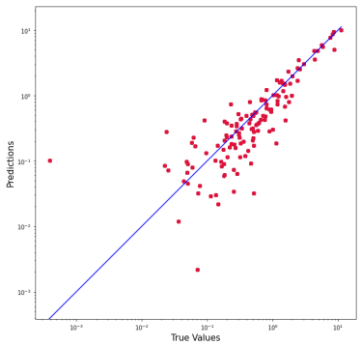
Regresyon Modelleri

1. **Lineer Regresyon:** Tahmin edilmek istenen değerler için tüm noktalara uzaklığı minimum olan bir doğrusal bir çizgi çekerek tahminleme yapar.
2. **Destek Vektörü Regresyonu:** Tahminleme yaparken bir doğru

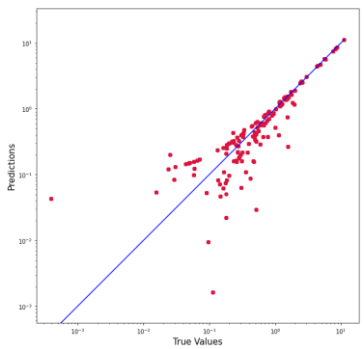
çizmektense bir aralık çizer ve o aralığı maksimum yapmaya çalışır.

3. **Karar Ağacı:** Bilgi kazanımının en yüksek olduğu özelliği bulma üzerine kurulu bir algoritmadır. Özellik çıkarımına veya normalizasyona gerek yoktur. Önemli özellikler zaten seçilmiş olacaktır.
4. **Yapay Sinir Ağları:** İnsan beyninin çalışma prensibinden esinlenerek geliştirilmiştir. Normalizasyon işlemi gereklidir. Girdi katmanı, gizli katman ve çıktı katmanından oluşur. Girdi katmanından gelen değerler ile ağırlıklarının nokta çarpımının bir aktivasyon fonksiyonundan çıktısı sonucunda bir sonraki katmandaki girdi elde edilmiş olur.
5. **Rassal Orman:** Eğitim aşamasında çok sayıda karar ağacı oluşturarak tahmin yapan bir toplu öğrenme yöntemidir.

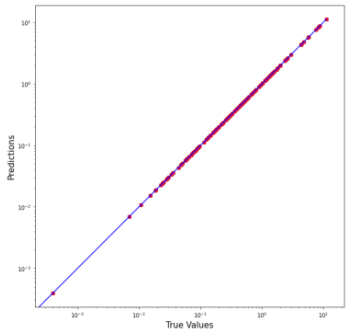
Veriler standardize edildikten sonra lineer regresyon algoritması uygulandığında aşağıdaki sonuçlar alınmıştır:



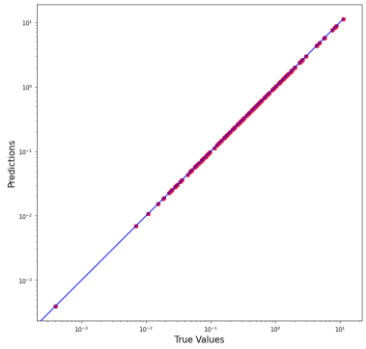
Destek vektör regresyonu ile alınan sonuçlar ise aşağıdaki gibidir:



Karar ağacı regresyonu uygulandığında ise sonuçlar aşağıdaki gibidir:



Rassal orman regresyonu ile alınan sonuçlar aşağıdaki gibidir:

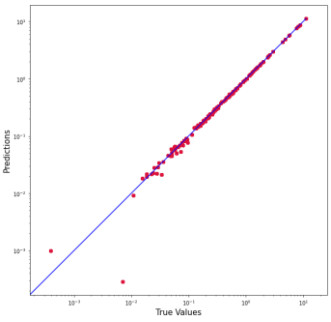


Yapay sinir ağı modeli aşağıdaki gibidir:

```
def build_model():
    model = Sequential()
    model.add(Dense(200, kernel_initializer = 'uniform', activation = 'relu', input_dim = X_train.shape[1]))
    model.add(Dense(100, kernel_initializer = 'uniform', activation='relu'))
    model.add(Dense(50, kernel_initializer = 'uniform', activation='relu'))
    model.add(Dense(1, activation='linear'))
    model.compile(loss="mse", optimizer="adam", metrics=["mse"])

    return model
```

Yapay sinir ağı kullanıldığında ise aşağıdaki gibi sonuç alınmıştır:



Sonuç

Değerlendirme puanı tahmininde en başarılı model K En Yakın Komşuluk olduğu gözlemlenmiş ama başarısı tatmin edici olmamıştır. Tüm modellerde verinin yoğun olarak 5 değerlendirme skoru içerdiğinden tahminleme yaparken de 5 deme eğiliminde olduğu gözlenmiştir. Kullanılan verilerin tahmin için yeterli olmadığı tespit edilmiştir.

Yapay veri arttırımı ile başarının arttırılabileceği düşünülmektedir.

Kitap satış miktarının tespitinde en iyi sonuç Karar Ağacı algoritmasından alınmıştır. Normalize edilmiş hatası(MSE) 1.01 * 10^(-6), denormalize edildikten sonraki hatası 556.8 çıktığı gözlemlenmiştir. En kötü sonuç ise Lineer Regresyondan alınmıştır. Karar ağaçlarında hiper parametre optimizasyonu ile daha iyi sonuçlar alınabileceği düşünülmektedir.

Referanslar

1. Wang, Bingkun, et al. "Review rating prediction based on user context and product context." Applied Sciences 8.10 (2018): 1849
2. Maghari, Alaa Mazen, Iman Ali Al-Najjar, and Said Jamil Al-Laqtah. "Books' Rating Prediction Using Just Neural Network." (2021).
3. Sharma, Satyendra Kumar, Swapnajit Chakraborti, and Tanaya Jha. "Analysis of book sales prediction at Amazon marketplace in India: a machine learning approach." Information Systems and e-Business Management 17.2 (2019): 261-284.