Journal of the Faculty of Engineering and Architecture of Gazi University 38:4 (2023) 2279-2289

Mühendislik Mimarlık Fakültesi Dergisi
Journal of The Faculty of Engineering and Architecture of Gazi University

Elektronik / Online ISSN : 1304 - 4915
Basılı / Printed ISSN : 1300 - 1884

# Analysis and price prediction of secondhand vehicles in Türkiye with big data and machine learning techniques

Burak Gülmez*, Sinem Kulluk
Department of Industrial Engineering, Faculty of Engineering, Erciyes University, 38039, Kayseri, Türkiye

**Highlights:**

- A dataset of secondhand vehicle prices in Türkiye is obtained
- Analyzes are made using big data analyses methods
- Price prediction is made on big data with machine learning algorithms

**Keywords:**

- Apache Spark
- Big data
- Regression algorithms

**Article Info:**

Research Article
Received: 14.08.2021
Accepted: 02.11.2022

**DOI:**

10.17341/gazimmfd.980840

**Correspondence:**

Author: Burak Gülmez
e-mail: burakgulmez@erciyes.edu.tr
phone: +90 505 829 5095

## Graphical/Tabular Abstract

In this study, first of all, data of secondhand vehicle market in Türkiye is obtained. Then, the collected data is converted into a secondhand vehicle prices dataset. The dataset contains many features of secondhand vehicles such as brand, model, year, fuel type, and gear type. The obtained big data is analyzed over Apache Spark and parallelization is applied while processing the data through Apache Spark. In this study, many analyzes are carried out related to the secondhand vehicles. In addition to the analyses, price prediction is performed by the characteristics of the vehicles using machine learning algorithms in Apache Spark. The random forest algorithm is outfperformed the other algorithms in price prediction. The graphical summary of the study is given in Figure A.

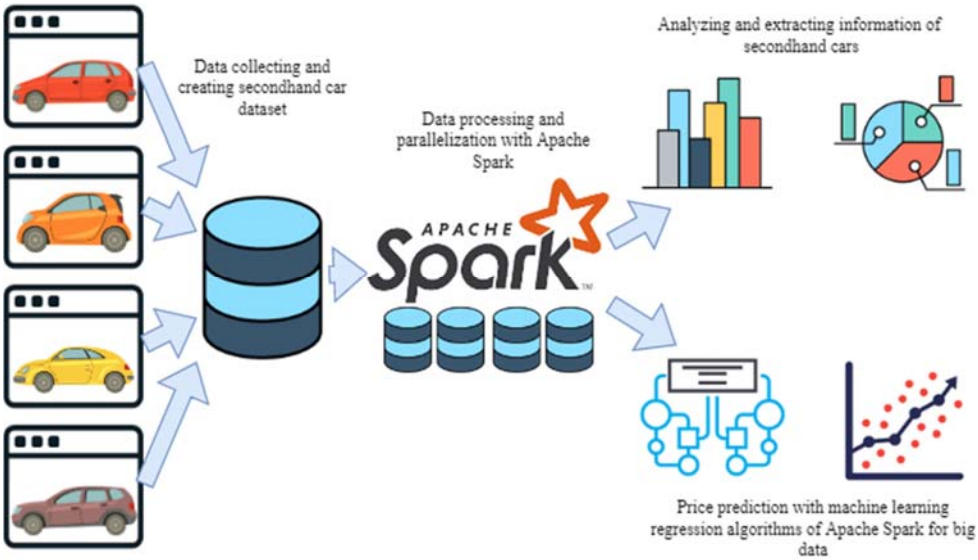



**Figure A.** Analysis and price prediction of secondhand vehicles with big data and machine learning techniques

**Purpose:** The purpose of this study is to analyze the secondhand vehicle dataset in Türkiye and to estimate the price of cars with Apache Spark and Apache Spark machine learning algorithms.

**Theory and Methods:** Information such as the most popular cars, most expensive cars, most popular models, the relationship between price and year are extracted by analyzing the dataset. Then linear regression, decision tree regression, random forest regression, GBT regression, and isotonic regression algorithms are used to estimate vehicle prices. Also, statistical analyzes are carried out to test whether the difference between the results obtained by the algorithms are statistically meaningful.

**Results:** The random forest algorithm is given the best results in price prediction with 21435.09 RMSE and 0.887 $R^2$. As a result of the statistical tests performed to control the significance of the differences between the RMSE and $R^2$ values obtained with the random forest and other algorithms, it is concluded that the random forest algorithm is statistically better than other algorithms.

**Conclusion:** The random forest algorithm is particularly successful in big data regression problems and can be used in this area. The success of the algorithm depends on performing training on multiple decision trees, its flexibility and strong hyperparameters.

Journal of the Faculty of Engineering and Architecture of Gazi University 38:4 (2023) 2279-2289

Mühendislik Mimarlık Fakültesi Dergisi
Journal of The Faculty of Engineering and Architecture of Gazi University

Elektronik / Online ISSN : 1304 - 4915
Basılı / Printed ISSN : 1300 - 1884

# Türkiye'de ikinci el araçların büyük veri ve makine öğrenme teknikleriyle analizi ve fiyat tahmini

Burak Gülmez*, Sinem Kulluk
Erciyes Üniversitesi, Mühendislik Fakültesi, Endüstri Mühendisliği Bölümü, 38039, Melikgazi, Kayseri, Türkiye

## Ö N E Ç I K A N L A R

- Türkiye'deki ikinci el araç fiyatları veri kümesi oluşturulmuştur
- Büyük veri yöntemleri kullanılarak analizler yapılmıştır
- Makine öğrenme algoritmaları ile büyük veri üzerinde fiyat tahmini yapılmıştır

Makale Bilgileri
Araştırma Makalesi
Geliş: 14.08.2021
Kabul: 02.11.2022

DOI:
10.17341/gazimmfd.980840

Anahtar Kelimeler:
Apache Spark,
büyük veri,
regresyon algoritmaları

ÖZ

Türkiye'de ikinci el araç piyasası her zaman hareketli olmuştur. İkinci el araç piyasasında marka, model, yakıt türü gibi özelliklerin ne kadar yoğunlukta olduğu, ne kadar fiyata etki ettiği gibi faktörler analiz edilerek, bu bilgiler kullanışlı hale getirilebilir. Araçların çeşitli özelliklerine göre fiyatları değişmektedir. Fiyatları tahmin edebilmek için makine öğrenme teknikleri kullanılabilir ve kullanıcıların araç satarken veya alırken fiyat belirlemelerine yardımcı olabilir. Fiyat tahmini, veri madenciliğinin bir görevi olan fonksiyon tahmini veya regresyon sınıfına girmektedir. İkinci el araç sayısı oldukça fazla olduğundan dolayı bu çalışmada analizler yapılırken büyük veri sistemleri kullanılmıştır. Apache Spark ve makine öğrenme kütüphanesi bunun için oldukça kullanışlıdır. Fiyat tahmini için doğrusal regresyon, karar ağacı regresyonu, rastgele orman regresyonu, GBT regresyonu, izotonik regresyon algoritmaları kullanılmıştır. Kullanılan algoritmalar ile araçların fiyat tahmini yapılmıştır ve en yüksek başarıyı 21435,09 RMSE ve 0,887 $R^2$ değerleriyle rastgele orman algoritması elde etmiştir. Rasgele orman algoritması ve diğer algoritmalarla elde edilen RMSE ve $R^2$ değerleri arasında anlamlı bir farklılık olup olmadığını kontrol için yapılan istatistiksel testler sonucunda, rasgele orman algoritması ile elde edilen sonuçların daha iyi olduğu sonucuna ulaşılmıştır. Rasgele orman algoritmasının daha iyi sonuçlar vermesinin nedeni, algoritmanın birden çok karar ağacı üzerinden eğitim gerçekleştirmesi, esnekliği ve güçlü hiper parametrelere sahip olmasıdır.

# Analysis and price prediction of secondhand vehicles in Türkiye with big data and machine learning techniques

## H I G H L I G H T S

- A dataset of secondhand vehicle prices in Türkiye is created
- Analyzes are made using big data methods
- Price prediction is made on big data with machine learning algorithms

Article Info
Research Article
Received: 14.08.2021
Accepted: 02.11.2022

DOI:
10.17341/gazimmfd.980840

Keywords:
Apache Spark,
big data,
regression algorithms

ABSTRACT

The secondhand vehicle market in Türkiye has always been active. In the secondhand vehicle market, information such as brand, model, and fuel type can be analyzed, and this information can be made useful. Prices vary according to the various features of the vehicles. Machine learning techniques can predict prices and help users set prices when selling or buying vehicles. Price prediction falls under regression. Since the number of secondhand vehicles is quite high, big data systems are used. Apache Spark and its machine learning library are quite useful for this. Linear regression, decision tree regression, random forest regression, GBT regression, and isotonic regression algorithms are used for price prediction. The random forest algorithm achieved the highest success for the price prediction with 21435.09 RMSE and 0.887 $R^2$ values. As a result of the statistical tests performed to check the significant difference between the RMSE and $R^2$ values obtained with the random forest algorithm and other algorithms, it is concluded that the results obtained with the random forest algorithm are statistically better than other algorithms. The random forest algorithm gives better results because the algorithm performs training over multiple decision trees, its flexibility, and strong hyperparameters.

*Sorumlu Yazar/Yazarlar / Corresponding Author/Authors : *burakgulmez@erciyes.edu.tr, skulluk@erciyes.edu.tr / Tel: +90 505 829 5095

## 1. Giriş (Introduction)

Günümüzde teknolojinin ve bilgisayarların gelişmesiyle birlikte veri toplamak daha kolay hale gelmiştir. Bu sayede veri depolama sistemlerinde geçmişe göre çok daha fazla veri tutulmaktadır ve depolanmaktadır. Veri madenciliği sayesinde tutulan verilerden anlamlı bilgiler çıkarılmakta ve çıkarılan bu bilgiler sayesinde karar alma süreçlerinde daha verimli hareket edilmektedir [1].

Büyük veri, klasik veri tabanı yönetim araçlarıyla ve veri işleme metotlarıyla işlenemeyecek kadar büyük miktardaki ve karmaşıklıktaki veri topluluklarına verilen addır. Veri miktarının çok olması ve kolay tutulması avantajlı bir durumdur, fakat büyük miktardaki veriyi yönetmek ve kullanmak zor olabilmektedir. Bu nedenle performans ve hız bakımından verimli metotlar ve araçlar kullanılmalıdır [2].

Verinin büyük veri olarak nitelendirilebilmesi için verinin bazı özelliklere sahip olması gerekmektedir. Bu özellikler çeşitlilik (variety), hacim (volume) ve hız (velocity) olarak sıralanabilir ve baş harfleri kullanılarak 3V olarak isimlendirilir. Bunlara ek olarak değişkenlik, karmaşıklık, doğruluk ve değer gibi özellikler de büyük veri özelliği olarak değerlendirilebilmektedir [2].

Veriler tek bir kaynaktan toplanmak zorunda değildir. Çeşitli internet siteleri, kayıt dosyaları, sosyal medya, e-postalar, sensörler gibi birden fazla farklı kaynaktan farklı yapılarda ve farklı miktarlarda veriler toplanabilir. Bu, verilerdeki çeşitliliği ifade etmektedir. Hacim, veri miktarındaki büyüklüğü temsil eder. Verinin türüne ve özelliklerine göre veri tabanları çok büyük yer kaplayabilmektedir. Popüler sosyal medya siteleri bir günde gigabayt ve terabaytlarca veri depolamaktadır [3]. Bunda video, resim gibi veri türleri kullanmaları büyük etkiye sahiptir. Hız, verinin hangi sıklıkta ve hangi miktarda geldiğiyle alakalı bir parametredir. Sadece veri gönderme aralığı değil, miktarı da önemlidir. Örneğin bir sensör topladığı verileri anlık olarak veya belli bir zaman biriktirip gönderebilmektedir. Değişkenlik, veri akışındaki tutarsızlıkları dikkate alır. Veriler, zamanla farklılaşma gösterebilir. Bu farklılaşmadan dolayı önce toplanan veriler ile sonra toplanan veriler arasında değişimler olabilir. Karmaşıklık, büyük verilerin farklı kaynaklardan üretildiğini ifade etmektedir. Farklı kaynaklardan gelen verileri birbirlerine bağlamak, eşleştirmek, temizlemek, düzenlemek gerekebilmektedir [4].

Büyük veri analizinde farklı sistemler kullanılmaktadır. Bu sistemlere örnek olarak Apache Hadoop, Apache Spark, Scikit-learn, Apache Mahout, Tensorflow, Keras verilebilir. Bu araçlar yardımıyla büyük veriler üzerinde işlemler kolaylıkla gerçekleştirilebilmektedir.

Apache Hadoop yazılım kütüphanesi, basit programlama modelleri kullanarak büyük veri kümelerinin bilgisayar kümeleri arasında dağıtılmasını sağlayan bir çerçevedir. Tek sunucudan binlerce makineye ölçeklenmek üzere tasarlanmıştır ve bunların her biri yerel hesaplama ve depolama sunmaktadır. Büyük veride özellikle Apache Hadoop için MapReduce yaklaşımı oldukça önemlidir. Bu yaklaşımda veri parçalanarak farklı yerlerde depolanır. Bazen aynı parçalar farklı yerlerde birden fazla kez tutulur. Bu sayede bir parçada bir problem çıkması durumunda, diğer parçalardan hatalı kısım telafi edilebilmektedir. Parçalar paralel şekilde işlenmekte ve işlenen her parçadan çıkan sonuç birleştirilmektedir. Böylece hız bakımından büyük bir avantaj sağlanmaktadır [3, 5, 6].

Apache Spark, büyük ölçekli veri işleme için hızlı ve genel bir motordur. Apache Hadoop ile yapılacak işlemleri çok daha hızlı bir şekilde gerçekleştirmektedir. Java, Scala, Python ve R programlama dillerini desteklemektedir. Bu nedenle kullanım çeşitliliği sunmaktadır. Apache Spark, Apache Hadoop'un daha yenisi ve gelişmişi olarak düşünülebilir. Apache Spark sayesinde büyük veri üzerinde sorgular ve makine öğrenme algoritmaları hızlı bir şekilde çalıştırılabilmektedir. Tek makinede çalışabildiği gibi birden fazla makinede de çalışabilme kabiliyeti vardır. RAM üzerinde işlemleri gerçekleştirdiği için oldukça hızlıdır [7].

Apache Mahout, Apache Hadoop üzerinde çalışmakta ve büyük veri üzerinde makine öğrenme algoritmalarını gerçekleştirmektedir. Algoritmaları çalıştırırken her seferinde diske yazdığı için Apache Spark'a göre daha yavaştır. Apache Spark işlemleri RAM üzerinde çalıştırmaktadır. Sckit-learn bir Python kütüphanesidir. Büyük veri ve paralel işlemler için özelleştirilmediği için yavaş kalabilir. Kullanılacağı takdirde Python içerisinde iyi bir paralelleştirme kodu yazılmalıdır [8].

Büyük veri analizinde makine öğrenme teknikleri sıklıkla kullanılmaktadır. Makine öğrenme teknikleri, makinelerin klasik kodlama dışına çıkarak bir insan gibi öğrenmeleri için tasarlanmış algoritmalardır. Makine öğrenme algoritmalarına girdi ile istenen çıktı verilir. Algoritma, bu girdileri kullanarak gerçekleştireceği değerlendirmeyi kendi içerisinde uygular [9, 10]. Literatürde makine öğrenimiyle ilgili çeşitli algoritmalar mevcuttur. Bu algoritmalar arasında en çok doğrusal regresyon, destek vektör makineleri, karar ağaçları, rastgele orman, yapay sinir ağları ve izotonik regresyon sıklıkla kullanılmaktadır. Bu makine öğrenme algoritmaları sayesinde veri üzerinden bilgi çıkarımı gerçekleştirilmektedir [11].

Bu çalışmada Türkiye'deki ikinci el araç fiyatları, Apache Spark ve makine öğrenme kütüphanesi kullanılarak tahmin edilmiştir. Bu doğrultuda ilk olarak, Türkiye'deki ikinci el araç verileri internet üzerindeki ilanlardan elde edilmiştir. Toplanan bu veri üzerinden ikinci el araçların en yaygın, an pahalı, en çok kullanılan yakıt türü gibi özellikleri istatistiksel olarak analiz edilmiş ve piyasa hakkında genel bilgilere ulaşılmıştır. Sonrasında fiyat tahmini için doğrusal regresyon, karar ağacı regresyonu, rastgele orman regresyonu, GBT regresyonu ve izotonik regresyon algoritmaları kullanılarak tahminler elde edilmiştir. Algoritmalarla elde edilen sonuçlar istatistiksel olarak karşılaştırılmıştır ve ikinci el araç fiyat tahmini için en uygun algoritma belirlenmiştir.

Makalenin izleyen bölümleri şu şekilde organize edilmiştir. İkinci bölümde literatürdeki benzer çalışmalara yer verilmiştir. Üçüncü bölümde, ikinci el araç verileri hakkında bilgiler ve veri analizi sonuçları verilmiştir. Dördüncü bölümde, araç fiyat tahmini için kullanılan algoritmalara değinilmiştir. Beşinci bölümde, ikinci el araç fiyat tahmini, makine öğrenme algoritmaları ile gerçekleştirilmiş ve elde edilen tahminler istatistiksel olarak karşılaştırılarak değerlendirilmiştir. Altıncı bölümde ise çalışmadan elde edilen sonuç ve önerilere yer verilmiştir.

## 2. Literatür İncelemesi (Literature Review)

Literatürde sınıflandırma alanında büyük veri sistemleri kullanılarak yapılan çok sayıda çalışma vardır. Bu çalışmalar arasında, Ahmed vd. [12] küçük bir diyabet verisi üzerinde diyabet tahmini için Apache Spark makine öğrenme algoritmalarını kullanılmış ve karşılaştırmışlardır. Ele aldıkları algoritmalardan lojistik regresyon algoritması; karar ağacı, rastgele orman, destek vektör makinesi ve Naive Bayes algoritmalarından daha iyi sonuç vermiştir.

Oo ve Thein [13] yaptıkları çalışmada, Spark yardımıyla büyük veriler için rastgele orman algoritmasından faydalanmışlardır. Bu algoritmayı nitelik indirgemeyle birlikte en iyi parametre değerleriyle kullanmışlardır. Algoritmayı beş büyük veri kümesi üzerinde test

etmişler ve Spark'ta bulunan rastgele orman algoritması ile karşılaştırmışlardır. Süre olarak daha kısa sürede sonuca ulaşmışlardır. Rio vd. [14] tarafından yapılan çalışmada dengesiz büyük miktardaki veri kümeleri için rastgele orman algoritması üzerinde çalışılmıştır. Rastgele orman algoritması bir sınıflandırma algoritmasıdır. Veri büyüklüğüyle baş edebilmek için çalışmada MapReduce tekniği kullanılmıştır. Farklı parametreler ile literatür verileri üzerinde analizler yapılmıştır. Kısa süreler içerisinde algoritmalar sonuçları verebilmiştir.

Sağlamlar [15] sınıflandırma için çok yüzlü konik sınıflandırıcıları test etmiştir. Sınıflandırma yaparken büyük veri ile uyumlu algoritmalar kullanmıştır. Önerdiği algoritma, diğer klasik algoritmalara göre daha başarılı sonuçlar vermiştir. Bu sayede önerdiği yöntemin büyük veri sınıflandırmasında etkili olduğunu göstermiştir.

Tao vd. [17] yüksek nitelik sayısı olan veriler için akıllı bir kümeleme algoritması geliştirmişlerdir. Normal kümeleme algoritmaları, yüksek nitelik sayısına sahip olan verilerde iyi sonuç vermemektedir, bunun sebebi bütün niteliklerin aynı ağırlıkta değerlendirilmesidir. Sürü algoritması ile k-ortalamalar algoritmasını hibritleyerek bu problemin üstesinden gelmişlerdir. Sürü algoritması ile kümelerin merkezleri ve verideki niteliklerin ağırlıkları optimize edilmiştir. Çok sayıda nitelik değerine sahip veri kümeleri üzerinde hem tek işlemci ile hem de Spark kullanarak paralel bir şekilde algoritmayı çalıştırmışlardır. Etkili sonuçlar elde etmişlerdir.

Alnafessah ve Casale [18], anomali saptaması için Apache Spark üzerinde yapay sinir ağları tabanlı teknikler denemişlerdir. Kayıt dosyaları ve işletim sisteminin izlenmesinden oluşan verileri analiz etmişlerdir ve bu sayede anomali saptamışlardır. Yöntemlerini karar ağacı, en yakın komşu ve destek vektör makinesi algoritmalarıyla karşılaştırmışlar ve diğer algoritmalardan daha doğru sonuçlar elde etmişlerdir. Yüzde 98'lere varan doğruluk değeri yakalamışlardır.

Kümeleme alanında yapılan büyük veri çalışmaları da mevcuttur. Lu [19] yaptığı çalışmada kümeleme performansını ve hızını artırmak için k-ortalamalar algoritması üzerinde geliştirme yapmıştır. Bu yeni algoritmayı Hadoop üzerinden hız performansını artırmak için test etmiştir. İlk olarak noktalar ile küme merkezleri arasında bir uzaklık aralığı belirlemiştir ve bu aralığı aşmayan noktaları o kümeye atamıştır. Dışarıda kalan noktaları ise en yakın kümeye atamıştır. Bu algoritmayı Hadoop ve Apache Mahout üzerinde test etmiş ve iyi sonuçlar elde etmiştir.

Cui vd. [20] kümeleme analizi üzerinde çalışmıştır. Kümeleme, en çok kullanılan veri işleme yöntemlerinden biridir. K-ortalamalar algoritması basitliği nedeniyle en çok kullanılan kümeleme algoritmasıdır. Son zamanlarda veri miktarı arttığı için araştırmacılar MapReduce'a yönelmeye başlamışlardır. MapReduce tekrarlanan iş, büyük veri okuma ve karıştırma süreleri nedeniyle yinelenen algoritmalar için uygun değildir. Büyük ölçekli verilerin K-ortalamalar kümeleme algoritması kullanılarak işlenmesi ve yinelemeye bağımlılığın ortadan kaldırılması ve yüksek performans elde etmek için MapReduce'da yeni bir işleme modeli önerilmiştir.

Birliktelik analizi için büyük veri üzerinde literatürde farklı çalışmalar vardır. Shang vd. [21] bulanık bir birliktelik analizi yöntemiyle şirketlerin finansal risklerini önceden saptamayı amaçlamışlardır. Bunun için ilk olarak korelasyon yardımıyla kullanılacak finansal gösterge sayısını 32'den 12'ye düşürüşümlerdir. Şirketlerin bu 12 göstergesini uzun dönemde inceleyip 4 farklı kümeye bölmüşlerdir ve bu kümeler üzerinden birliktelik analizi uygulamışlardır. Moens vd. [22] yaptıkları çalışmada birliktelik analizi veya market sepet analizi için kendi geliştirdikleri iki algoritmayı test etmişlerdir. MapReduce tekniği sayesinde bu iki algoritma büyük veri üzerinde test edilmiştir. Algoritmalar başarı ve süre bakımından karşılaştırılmıştır.

Zhang vd. [23] DFIMA (distributed frequent itemset mining algorithm) adında birliktelik analizi üzerine bir algoritma önermişlerdir. Bu algoritma Apache Spark üzerinde çalıştırılmıştır. DFIMA ve FP-Growth algoritmaları 4 MB, 15 MB ve 2 GB boyutlarındaki veri kümeleriyle test edilmiştir. Çalışılan algoritma FP-Growth algoritmasından daha iyi sonuçlar vermiştir.

Nodarakis vd. [24] Twitter üzerinde büyük ölçekli duygu analizi (sentiment analysis) çalışmışlardır. Klasik yöntemlerle Twitter paylaşımlarının çok az bir kısmı analiz edilebilmektedir. Bu durumu geliştirmek ve daha yüksek sayıda analiz yapabilmek için Apache Spark üzerinde çalışan bir metot geliştirmişlerdir. kNN (k en yakın komşu) algoritması üzerinde çalışmışlar ve yüksek performanslar elde etmişlerdir.

Alaoui vd. [25] sosyal medya verilerini analiz etmişlerdir. 2016 yılındaki ABD başkanlık seçimleri ile alakalı yazılar üzerinden metin madenciliği yapmışlardır. Bunun için Kafka, Apache Spark ve Apache Hadoop kullanmışlardır. Twitter verileri üzerinden analiz yapmışlardır. Analizlerinde Donald Trump ve Hillary Clinton'u karşılaştırmışlardır. Trump, Clinton'dan daha iyi bir skor elde etmiştir.

Hasan vd. [26] Twitter'da akan veri için sınıflandırma ve kümeleme üzerinde çalışmışlardır. Twitter verisi dışında Higgs verisi üzerinde de çalışmışlardır. Algoritma olarak gelişmiş bulanık C ortalama algoritması ve parçacık sürü algoritması kullanmışlardır. Doğruluk, F skor, kesinlik, duyarlılık gibi ölçütler üzerinde destek vektör makinesi ve anti-bayes sınıflandırma algoritmalarından daha iyi sonuçlar elde etmişlerdir.

Altıntaş vd. [27] metin madenciliği üzerinde çalışmışlardır. Öncelikle internet ortamındaki sağlık ile alakalı paylaşımları toplamışlardır. Çok fazla sayıda paylaşım olduğu için çok büyük miktarda veri elde etmişlerdir. Büyük veri üzerinde geliştirdikleri yöntem sayesinde kanser ile alakalı paylaşımları süzmüşlerdir. Yaptıkları tutarlılık testlerinde iyi sonuçlar elde ettiklerini görmüşlerdir.

Regresyon alanında da büyük veri çalışmaları mevcuttur. Syed vd. [28] Londra'nın enerji harcama hesaplaması tahmini için bir çalışma yapmışlardır. Kısa, orta ve uzun dönemler için tahminler elde etmişlerdir. Apache Spark ve Hadoop kullanmışlardır. Spark'ın makine öğrenme kütüphanesi olan Mllib'ten de yararlanmışlardır. Bu makine öğrenme kütüphaneleri içerisindeki regresyon, rastgele orman, karar ağacı gibi algoritmaları denemişlerdir. Çıkan sonuçlarda Spark sayesinde oldukça hızlı ve başarısı yüksek değerler elde etmişlerdir.

Taşyürek ve Çelik [29] çalışmalarında konum temelli bir regresyon yöntemi olan coğrafi ağırlıklı regresyon metodunu kullanmışlardır. Bu yöntem ile mekânsal ve zamansal olarak regresyon yapılabilmektedir. Fakat yöntem, hız ve performans bakımından oldukça etkisiz kalabilmektedir. Bu yüzden yöntemi geliştirmişler ve büyük veri üzerinde kolayca çalışabilen bir hale dönüştürmüşlerdir. Bu sayede hem hız bakımından hem de performans bakımından ilerleme kaydetmişlerdir. Yöntemi, büyük veriler için kullanılabilir bir versiyona dönüştürmüşlerdir.

Arslan ve Aslan [30] yapay arı koloni (ABC) algoritmasının gelişmiş bir hali olan kafes temelli ABC (LBABC) algoritmasını geliştirmişlerdir. Bu algoritmayı büyük veriler için uygun hale dönüştürmüşlerdir. Önerdikleri algoritmayı bazı büyük veri gürültü minimizasyonu problemleri üzerinde test ermiş ve diğer bazı

metasezgisellerle karşılaştırmışlardır. Önerdikleri algoritma ile daha iyi sonuçlar elde etmişlerdir. Xu vd. [31] yaptıkları çalışmada rüzgar hızını tahmin etmeye çalışmışlardır. Bunun için Çin'de her 3 dakikada bir rüzgâr hızı kayıtlarını içeren bir veri kümesi kullanmışlar ve bu veriler üzerinden zaman serisi çalışması yapmışlardır. Yapay sinir ağlarından yararlandıkları bu çalışmalarını Spark üzerinden paralel ve dağıtık bir biçimde yaptıkları için tek bir bilgisayarda çalışmaya göre daha hızlı sonuçlar almışlardır.

Manogaran ve Lopez [32] büyük veri analizini iklim değişikliğini gözlemlemek için kullanmışlardır. Çok büyük miktardaki iklim verileri eski yöntemlerle tutulamamaktadır. Yaptıkları çalışmada Apache Hadoop altyapısıyla MapReduce algoritmasını kullanmışlardır. İklimsel parametrelerin mevsimsel olarak ölçümleri yapılmış ve ortalamalar hesaplanmıştır. Kümülatif ortalamalar üzerinden grafikler takip edilmiştir.

Xu vd. [31] rüzgar hızını ölçmek için Apache Spark üzerinde çalışan bir yapı kurmuşlardır. Enerji üretimi için kritik bir durumda olan rüzgâr hızı ölçümleri eski yöntemlerde çok iyi tahminler vermemektedir. Bunun için dağıtılmış bir veri işleme sistemi kurmuşlardır. Hem hız bakımından hem de tahmin doğruluğu bakımından daha başarılı sonuçlar elde etmişlerdir. Bu çalışmada Türkiye'deki ikinci el araç pazarı için öncelikle, büyük veri sistemlerinde kullanılan Apache Spark ile analizler yapılmıştır. Bu analizler marka, model, vites türü, yakıt türü gibi birçok etkeni içermektedir. Sonrasında ikinci el araç fiyat tahmini gerçekleştirilmiştir. Fiyat tahmini için büyük veri analizinde sıkça kullanılan Apache Spark üzerinden doğrusal regresyon, izotonik regresyon, karar ağacı regresyonu, rastgele orman regresyonu ve GBT regresyon olmak üzere beş farklı algoritma kullanılmıştır. Bu algoritmalar sayesinde gerçek değerlere çok yakın fiyat tahminleri elde edilmiştir.

## 3. Araç Fiyatları Verisi Analizi
**(Automobile Prices Data Analysis)**

Türkiye'de ikinci el araç piyasası oldukça hareketlidir. Çok fazla sayıda araç ikinci el pazarında alınıp satılmaktadır. Araçların fiyatına etki eden faktörler marka, model, model yılı, yakıt türü, vites türü, motor hacmi, motor gücü, kilometre, hasar durumu, takas durumu, satıcı kişinin durumudur. Bu faktörlere bağlı olarak fiyat tahmini yapılmaktadır.

İkinci el araç verisi, 2019 yılında internet üzerinde araçlarını satmak isteyen kişilerin koydukları ilanlardan toplanmıştır. Araç verisi toplam 120000 örnekten oluşmaktadır ve her örnek için 11 özellik değeri ve bir fiyat değeri mevcuttur. Veri kümesinden, fiyatı 1000000 TL (bir milyon TL) üzeri ve 1000 TL (bin TL) altında olan örnekler çıkartılmıştır. Bunun sebebi, fiyat giren kişilerin fiyat yazarken fazla veya az sıfır koymalarıdır. Örneğin araç fiyatı 75000 TL (yetmiş beş bin TL) ise araç sahibi fiyatı girerken kuruş cinsinden düşünerek üç sıfır fazla yazmaktadır ve 75000000 TL (yetmiş beş milyon TL) olarak veri girişi yapmaktadır. Buna benzer şekilde 75 TL (yetmiş beş TL) olarak da veri girişi yapanlar mevcuttur. Bu durumların veri kümesini olumsuz etkilememesi için bu aykırı değerleri veri kümesinden çıkararak düzenleme yapılmıştır. Veri kümesine ait beş adet örnek, örnek olarak Tablo 1'de verilmiştir.

İkinci araç veri kümesinden çeşitli bilgiler elde edilmiştir. Bunlardan bazıları en çok listelenen marka, en çok listelenen model, araçların yakıt dağılımı, en pahalı veya ucuz modeller gibi bilgilerdir.

### *3.1. Marka Analizi (Analysis of Brand)*

Veri kümesinde toplam 52 adet marka mevcuttur. Bu markalardan en çok bulunanı Renault markasıdır. Marka, ilanlarda kaç adet o markadan bulunduğu, o markanın ortalama fiyatına ait 10 adet örnek Tablo 2'de görüldüğü gibidir.

**Tablo 2**. Markaların sayı, ortalama fiyat bilgisi
(number and average price of the brands)

| Model | Sayı | Ortalama fiyat |
|---|---|---|
| Renault | 16872 | 85166,22 |
| Volkswagen | 13820 | 120638,78 |
| Opel | 10528 | 77014,83 |
| Fiat | 9699 | 62803,05 |
| Ford | 9175 | 93674,51 |
| Hyundai | 7246 | 84018,49 |
| Toyota | 5631 | 98229,87 |
| Peugeot | 4272 | 64326,34 |
| Honda | 3393 | 110280,61 |
| BMW | 2705 | 158483,67 |

### *3.2. Model analizi (Analysis of Model)*

Marka analizine benzer olarak model analizi de yapılmıştır. En çok ilanda bulunan modellerden 10 tanesi Tablo 3'te verilmişti. En çok ilanda bulunan marka Renault olmasına rağmen, en çok ilanda bulunan ilk iki model Ford ve Opel markalarının modelleri çıkmıştır. Fakat Renault dört farklı modelle bu listede yer almıştır. Ayrıca Renault marka modeller listede diğerlerine göre ortalama olarak daha düşük fiyattadır.

### *3.3. Yakıt Türü Analizi (Analysis of Fuel Type)*

Yakıt türü incelendiğinde en çok dizel araç bulunmaktadır. Dizel araçları sırasıyla LPG&benzin, benzin ve hibrit araçlar takip etmektedir. Fiyat olarak ise en yüksek fiyatlı araçlar hibrit araçlardır. Sonrasında dizel, benzin ve LPG&Benzin araçlar gelmektedir. Yakıt türü için sayı, ortalama fiyat bilgisi Tablo 4'te verilmiştir.

**Tablo 1**. İkinci el araç veri kümesi (Secondhand vehicle dataset)

| Marka | Alfa Romeo | Alfa Romeo | Alfa Romeo | Alfa Romeo | Alfa Romeo |
|---|---|---|---|---|---|
| Model | Giulietta | 159 | Giulietta | Giulietta | Giulietta |
| Yıl | 2016 | 2006 | 2011 | 2011 | 2014 |
| Yakıt türü | Benzin | Dizel | Benzin | Dizel | Dizel |
| Vites | Otomatik | Düz | Düz | Düz | Düz |
| Motor hacmi | 1368 | 1910 | 1368 | 1598 | 1598 |
| Güç | 170 | 150 | 120 | 105 | 105 |
| Km | 50000 | 297000 | 110000 | 127500 | 82000 |
| Hasar | Boyalı | Değişenli | Belirtilmemiş | Boyalı | Tamamı orijinal |
| Takas | Takasa uygun değil | Takasa uygun değil | Takasa uygun değil | Takasa uygun değil | Takasa uygun |
| Satıcı | Sahibinden | Sahibinden | Sahibinden | Sahibinden | Galeriden |
| Fiyat | 175000 | 78000 | 119999 | 122000 | 124900 |

**Tablo 3.** Modellerin sayı, ortalama fiyat bilgisi
(Number and average price of the models)

| Marka | Model | Sayı | Ortalama fiyat |
|---|---|---|---|
| Ford | Focus | 6207 | 106607,30 |
| Opel | Astra | 6198 | 89137,79 |
| Renault | Clio | 5315 | 93585,14 |
| Toyota | Corolla | 4085 | 101725,60 |
| Volkswagen | Passat | 3401 | 143580,80 |
| Volkswagen | Golf | 3377 | 130411,30 |
| Renault | Fluence | 3082 | 99020,49 |
| Renault | Megane | 3052 | 79311,62 |
| Renault | Symbol | 3004 | 76592,84 |
| Honda | Civic | 2872 | 116631,90 |

**Tablo 4**. Yakıt türü için sayı, ortalama fiyat bilgisi
(Number and average price for fuel type)

| Yakıt türü | Sayı | Ortalama fiyat |
|---|---|---|
| Dizel | 59554 | 109261,70 |
| LPG&Benzin | 32725 | 60682,87 |
| Benzin | 29285 | 108458,40 |
| Hibrit | 9 | 155444,40 |

*3.4. Model Yılı Analizi (Analysis of Model Year)*

Model yıllarına bakıldığında Tablo 5'e göre en çok aracın 2012 model araçlar olduğu görülmektedir. Model yıllarından dolayı sayılarda çok büyük farklar oluşmamıştır. 2011 – 2017 yıllarına ait modeller en çok ilanda bulunan modellerdir.

**Tablo 5.** Yıllara göre sayı, ortalama fiyat bilgisi
(Number and average price for years)

| Yıl | Sayı | Ortalama fiyat |
|---|---|---|
| 2012 | 9666 | 106587,92 |
| 2013 | 9375 | 109872,11 |
| 2014 | 8239 | 114829,04 |
| 2015 | 8063 | 130615,64 |
| 2011 | 7995 | 85607,02 |
| 2016 | 6693 | 142806,58 |
| 2017 | 6595 | 146875,18 |
| 2018 | 4716 | 166765,33 |
| 2010 | 4701 | 84945,59 |
| 2006 | 4170 | 59481,77 |
| 2004 | 4105 | 52660,10 |

Yıllara göre araç değerleri artmaktadır. Bu artış Şekil 1'de görülebilir.

*3.5. Vites Türü Analizi (Analysis of Gear Type)*

Vites türüne göre araçlar incelenecek olunursa en çok bulunan vites türü düz vitestir. Düz vitesi yarı otomatik ve otomatik takip etmektedir. Düz vitesli araçlar diğerlerine göre oldukça fazladır ve fiyat olarak ise diğerlerinden ucuzdur. En pahalı araçlar ise otomatik vitesli araçlardır. Vites durumuna göre veri kümesindeki araç sayısı ve ortalama fiyat bilgisi Tablo 6'da gösterilmiştir.

**Tablo 6.** Vites türüne göre sayı, fiyat analizi
(Number and average price for gear type)

| Vites | Sayı | Ortalama fiyat |
|---|---|---|
| Duz | 76524 | 73391,74 |
| YariOtomatik | 26415 | 136440,30 |
| Otomatik | 18634 | 148032,10 |

*3.6. Takas Durumu Analizi (Analysis of Swap)*

Tablo 7'deki bilgilere göre araçların takas durumlarına bakılacak olunursa, takasa uygun olmayan araçlar daha fazladır. Fakat takasa uygun araç sayısı da azımsanamayacak sayıdadır. Bu durum, bu sektörden ticari olarak para kazanan insanların çokluğunu göstermektedir. Ayrıca takasa uygun olarak satılan araçların fiyat ortalaması yaklaşık 5000 TL daha fazladır. Bu da ikinci el araç alım satımı yapan insanlar veya galerilerin ortalama olarak 5000 TL daha pahalıya araç sattığını göstermektedir.

**Tablo 7.** Takas durumunun sayı ve fiyat analizi
(Number and average price for swap type)

| Takas durumu | Sayı | Ortalama fiyat |
|---|---|---|
| TakasaUygunDegil | 69406 | 92864,49 |
| TakasaUygun | 52167 | 97738,26 |

*3.7. Satıcı Türü Analizi (Analysis of Seller Type)*

Satıcı türlerine göre veri kümesi incelenirse, en çok sahibinden araç satılmaktadır. Sahibinden satılan araç sayısının yaklaşık olarak yarısı kadar ise galeriden araç satılmaktadır. Bu durum Tablo 8'den görülebilir.

**Tablo 8.** Satıcı türünün sayı ve fiyat analizi
(Number and average price for seller type)

| Satıcı | Sayı | Ortalama fiyat |
|---|---|---|
| Sahibinden | 79345 | 88998,06 |
| Galeriden | 39971 | 105510,40 |
| YetkiliBayiden | 2253 | 140961,80 |
| RentaCar | 4 | 153562,50 |

## 4. İkinci El Araç Fiyat Tahmini İçin Kullanılan Algoritmalar
**(The Algorithms Used for Secondhand Vehicle Price Prediction)**

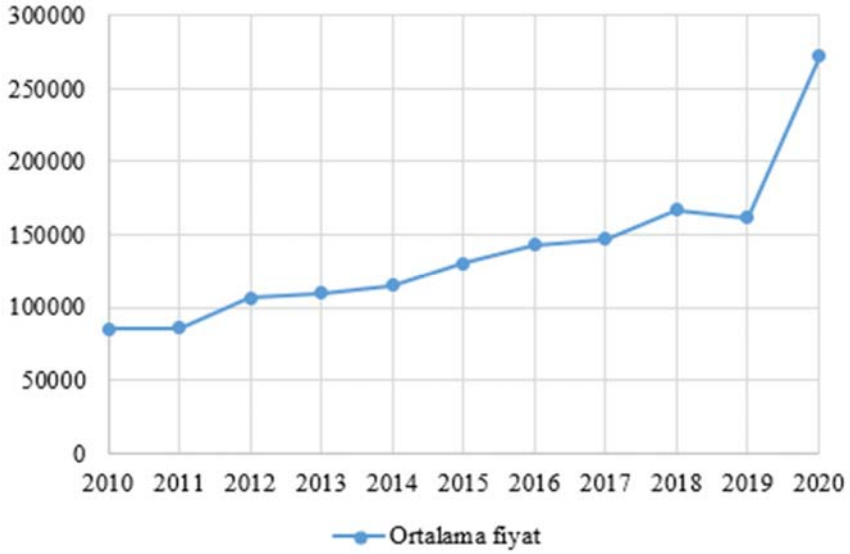
Araç verilerinin analizinin yanı sıra çalışmada araç fiyat tahmini de gerçekleştirilmiştir. Bu doğrultuda makine öğrenme metotlarından fonksiyon tahminleme (regresyon) kullanılmıştır. Bu algoritmaların kullanılabilmesi için veri kümesindeki metin (string) şeklindeki verilerin öncelikle ön işlem olarak sayısal değerlere çevrilmesi gerekmektedir. Bunun için metinleri kategoriye çeviren ve onlara sırasıyla tam sayı değer veren dönüştürücü (StringIndexer) kullanılmıştır. Bu dönüşümden sonra algoritmalar fonksiyon tahminleme amacıyla kolaylıkla kullanılabilecektir.

Kullanılan algoritmalar doğrusal regresyon (linear regression), karar ağacı regresyonu (decision tree regression), rastgele orman regresyonu (random forest regression), GBT regresyon (gradient-boosted tree regression) ve izotonik regresyon (isotonic regression) algoritmalarıdır. Bu algoritmalar çok sık kullanılan ve iyi sonuçlar veren makine öğrenme algoritmalarıdır.

Algoritmaların karşılaştırılması için ortalama kareli hata kökü RMSE (root mean square error) ve $R^2$ metotları kullanılmıştır. RMSE formülü Eş. 1'de, $R^2$ formülü ise Eş. 2'de verilmiştir. Bu eşitliklerde $n$, örnek sayısını temsil etmektedir. $y_i$, tahmin değerini; $y$ ise gerçek değeri ifade etmektedir. $\hat{y}$, gerçek değerlerin ortalamasını temsil etmektedir.

Makine öğrenme algoritmaları ile tahmin üretmek için veriyi eğitim ve test kümelerine ayırmak gerekmektedir. Bu amaçla, k-katlı çapraz geçerlilik (K folds cross validation) veya yüzde bölümleme (percentage split) kullanılabilir. Bu çalışmada veri, eğitim ve test kümelerine %70-%30 eğitim-test oranlarında yüzde bölümleme ile ayrılmıştır.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - y)^2} \quad (1)$$



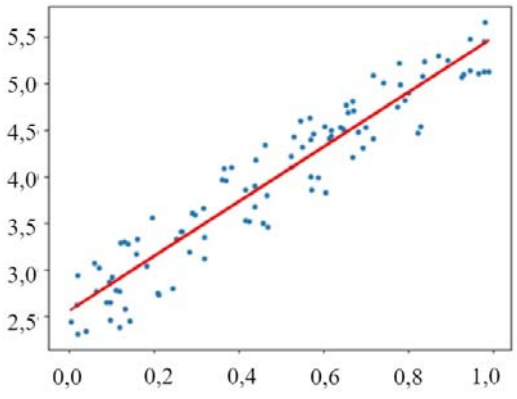

**Şekil 1.** Yıllara göre ortalama fiyat değerleri (Average price for years)

$$R^2 = 1 - \frac{\sum_i^n (y_i - \hat{y})^2}{\sum_i^n (y_i - \bar{y})^2} \quad (2)$$

*4.1. Doğrusal Regresyon (Linear Regression)*

Doğrusal regresyonda, bağımsız değişkenler ile hedeflenen değişken arasındaki ilişkiyi bulmak hedeflenir. Hangi değişkenin ne oranda etkili olduğu saptanır. Değişkenler ile hedeflenen değişken arasında doğrusal bir fonksiyon elde edilmeye çalışılır. Doğrusal regresyon formülü Eş. 3'teki gibidir [33, 34]. Bu eşitlikte *y*, elde edilecek olan tahmin değerini göstermektedir. $a_i$, *i.* değişkenin katsayısını temsil etmektedir. $x_i$ ise *i*. sıradaki değişkeni temsil etmektedir. Son olarak *b* değeri ise sabit bir değeri (kesişim) ifade etmektedir. Doğrusal regresyon örneği Şekil 2'de görüldüğü gibidir.

$$y = a_1 * x_1 + a_2 * x_2 + \cdots + b \quad (3)$$




**Şekil 2**. Doğrusal regresyon (Linear regression) [35]

*4.2. Karar Ağacı Regresyonu (Decision Tree Regression)*

Karar ağacı yönteminde bir ağaç oluşturulur ve değişkenlerin aldıkları değerlere göre ağaç dallanır. Ağaç dallandıkça ve derinleştikçe hedef değerlere yakın değerler elde edilir. Ağacın nasıl dallanacağı, hangi değişkenler üzerinden dallanacağı gibi kararlar sonuç için önemli faktörlerdir [36,37]. Bir karar ağacı örneği Şekil 3'te görüldüğü gibidir.

*4.3. GBT Regresyonu (Gradient Tree Regression)*

Bir karar ağacı yöntemidir. Öncelikle bir karar ağacı oluşturulur ve bu karar ağacından çıkan tahminler ile gerçek değerler arasındaki farklar hata olarak hesaplanır. Hesaplanan bu hatalardan yeni karar ağaçları oluşturulur ve sisteme eklenir. Bu adım benzer şekilde devam eder ve her ağaç eklendiğinde bir katsayı ile hataları gitgide düşürür [39]. Bir GBT regresyon örneği Şekil 4'teki gibidir.

*4.4. Rastgele Orman Regresyonu (Random Forest Regression)*

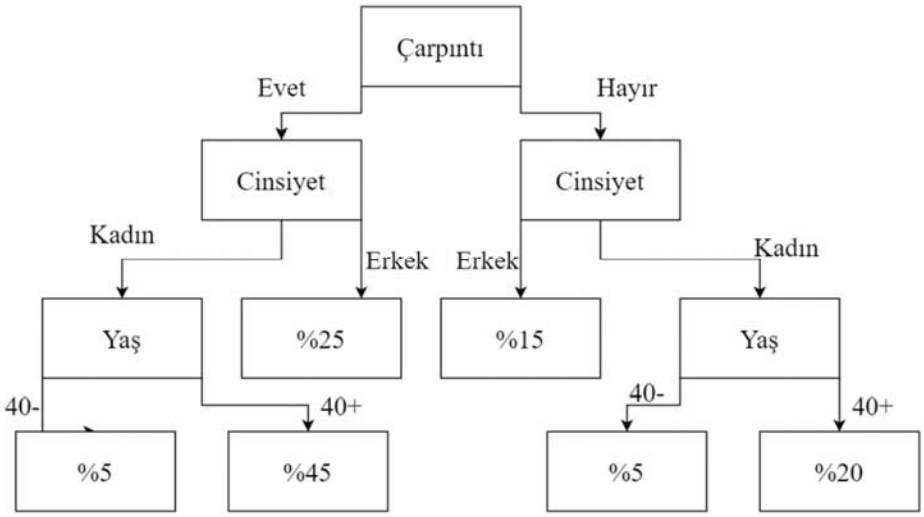
Bir karar ağacı yöntemidir. Veriyi rastgele şekilde küçük parçalara böler. Bölünen bu küçük parçalar üzerinden karar ağacı algoritması çalıştırılır ve ağaç sayısı kadar tahmin elde edilir. Test aşamasında ise bütün ağaçlardan gelen bilgiler birleştirilir ve bir değer elde edilir. Elde edilen değer rastgele orman yönteminin sonucudur [41]. Bir rastgele orman örneği Şekil 5'teki gibidir.

*4.5. İzotonik Regresyon (Isotonic Regression)*

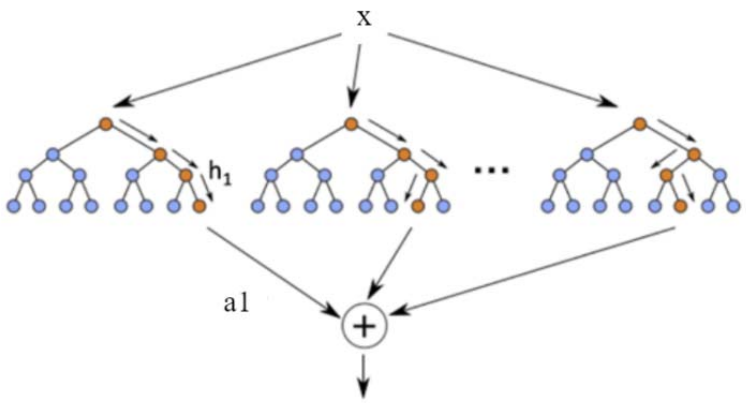
İzotonik regresyon, doğrusal regresyona oldukça benzemektedir. Doğrusal regresyondan farklı olarak tek bir formüle sahip değildir. Farklı aralıklarda farklı regresyon formüllerine sahiptir. Çözüm uzayını böler, parçalara ayırır ve her parça için farklı regresyon formülü uygular [42]. Bir izotonik regresyon örneği Şekil 6'da görüldüğü gibidir.

## 5. Elde Edilen Sonuçlar ve Tartışmalar (Obtained Results and Discussions)

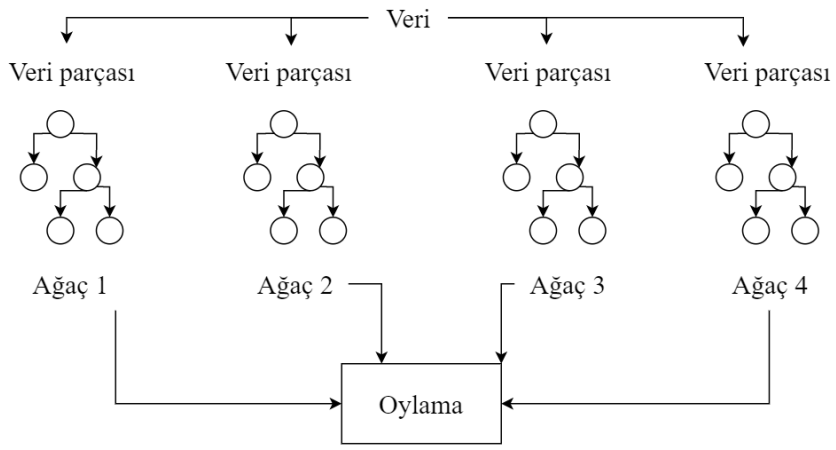
İkinci el araba verisi üzerinde doğrusal regresyon, karar ağacı, GBT, rastgele orman ve izotonik regresyon algoritmaları test edilmiştir. Veri kümesi %70 eğitim verisi ve %30 test verisi olarak ayrılmıştır. Performans ölçütü olarak RMSE ve $R^2$ ölçütleri kullanılmıştır. Algoritmalar 10 kez çalıştırılıp, test edilmiştir. Tablo 9'da algoritmaların RMSE karşılaştırması yapılmıştır. Tablo 10'da ise $R^2$ karşılaştırmalarına yer verilmiştir.



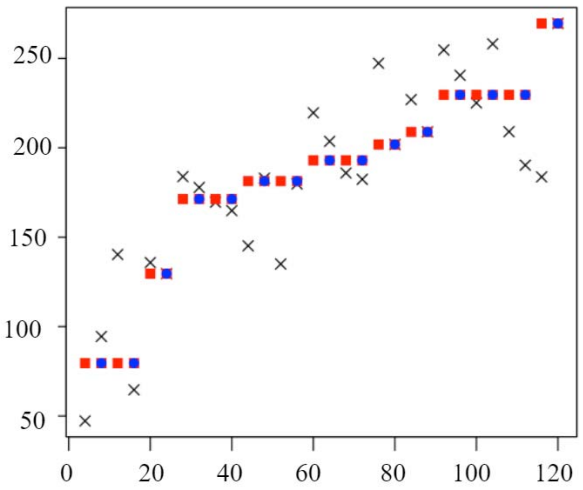

**Şekil 3.** Karar ağacı (Decision tree) [38]




**Şekil 4.** GBT regresyon (GBT regression) [40]




**Şekil 5.** Rastgele orman (Random forest)




**Şekil 6.** İzotonik regresyon (Isotonic regression) [43]

Algoritmalar 10 kez çalıştırılıp test edilmiştir. En düşük RMSE değerine sahip ve en yüksek $R^2$ değerine sahip algoritma rastgele orman algoritması çıkmıştır. Rastgele orman algoritmasını GBT ve karar ağacı algoritmaları takip etmektedir. İzotonik regresyon algoritması oldukça kötü sonuçlar vermiştir. Standart sapmalara bakıldığında en düşük değerleri RMSE için doğrusal regresyon yöntemi, $R^2$ için ise izotonik regresyon yöntemi vermiştir.

Sonuçları istatistiksel olarak karşılaştırmak için 2-örnek-t-test kullanılması planlanmıştır. 2-örnek-t-test kullanabilmek için öncelikle verilerin normal dağılıma uyup uymadığı kontrol edilmelidir. Bu kontrol için Kolmogorov-Smirnov testi gerçekleştirilmiştir. Kolmogorov-Smirnov test sonuçları RMSE için Tablo 11, $R^2$ için ise Tablo 12'de gösterilmiştir. Çıkan sonuçlara göre p değerlerinin tamamı 0,05'ten büyüktür. Dolayısıyla bütün değerler normal dağılıma uymaktadır. Dolayısıyla sonuçları istatistiksel olarak karşılaştırmak için 2-örnek-t-testi kullanılabilir.

Sonuçlar istatistiksel olarak karşılaştırıldığında ise Tablo 13 ve Tablo 14 elde edilmiştir. Tablolarda 2-örnek-t-testi için p değerleri gösterilmektedir. Eğer p değerleri 0,05'ten küçük ise farklar istatistiksel olarak anlamlıdır, aksi takdirde farklar anlamlı değildir.

İstatistiksel analiz sonuçlarına göre, GBT regresyon ile karar ağacı regresyon algoritmaları arasındaki fark istatistiksel olarak anlamlı çıkmamıştır. Bunun dışındaki bütün algoritmalar arasındaki farklar anlamlı farklardır, rastgele orman regresyon algoritması ile elde edilen sonuçlar istatistiksel olarak diğer algoritmalarla elde edilen

**Tablo 9.** Algoritmaların RMSE karşılaştırması (Comparison of the algorithms for RMSE)

| RMSE | Doğrusal Regresyon | Karar ağacı regresyon | GBT regresyon | Rastgele orman regresyon | İzotonik regresyon |
|---|---|---|---|---|---|
| 1 | 37921,80 | 32744,09 | 31417,59 | 22330,26 | 63861,98 |
| 2 | 37841,17 | 32895,29 | 32639,94 | 21103,40 | 62466,26 |
| 3 | 37389,27 | 33801,95 | 33913,53 | 20769,75 | 64112,21 |
| 4 | 37211,65 | 31987,28 | 30275,89 | 20850,38 | 63730,68 |
| 5 | 36958,79 | 31080,91 | 30103,47 | 20835,27 | 63385,51 |
| 6 | 38869,74 | 29808,31 | 25136,80 | 21892,78 | 64580,41 |
| 7 | 37153,44 | 31018,33 | 28848,29 | 20499,54 | 63366,62 |
| 8 | 38246,49 | 31974,17 | 28784,21 | 21486,44 | 65022,35 |
| 9 | 37844,89 | 34109,44 | 32777,43 | 21941,97 | 64801,35 |
| 10 | 37226,09 | 34040,60 | 33249,47 | 22641,08 | 63904,47 |
| Ortalama | 37666,33 | 32346,04 | 30714,66 | *21435,09* | 63923,18 |
| Standart sapma | *561,56* | 1364,97 | 2529,87 | 696,74 | 721,78 |

**Tablo 10.** Algoritmaların $R^2$ karşılaştırması (Comparison of the algorithms for $R^2$)

| $R^2$ | Doğrusal Regresyon | Karar ağacı regresyon | GBT regresyon | Rastgele orman regresyon | İzotonik regresyon |
|---|---|---|---|---|---|
| 1 | 0,645 | 0,735 | 0,756 | 0,877 | 0,007 |
| 2 | 0,632 | 0,722 | 0,726 | 0,886 | 0,003 |
| 3 | 0,654 | 0,717 | 0,715 | 0,893 | 0,017 |
| 4 | 0,660 | 0,749 | 0,775 | 0,893 | 0,003 |
| 5 | 0,658 | 0,758 | 0,773 | 0,891 | 0,007 |
| 6 | 0,637 | 0,786 | 0,848 | 0,885 | 0,003 |
| 7 | 0,655 | 0,759 | 0,792 | 0,895 | 0,004 |
| 8 | 0,653 | 0,757 | 0,803 | 0,890 | 0,004 |
| 9 | 0,656 | 0,721 | 0,742 | 0,884 | 0,008 |
| 10 | 0,655 | 0,712 | 0,725 | 0,872 | 0,017 |
| Ortalama | 0,650 | 0,742 | 0,766 | *0,887* | 0,007 |
| Standart sapma | 0,009 | 0,023 | 0,039 | 0,007 | *0,005* |

**Tablo 11.** RMSE için Kolmogorov-Smirnov testi (Kolmogorov-Smirnov test for RMSE)

| RMSE | Doğrusal Regresyon | Karar ağacı regresyon | GBT regresyon | Rastgele orman regresyon | İzotonik regresyon |
|---|---|---|---|---|---|
| Kolmogorov-Smirnov test değeri | 0,189 | 0,157 | 0,178 | 0,199 | 0,120 |
| p değeri | 0,804 | 0,935 | 0,862 | 0,752 | 0,995 |

**Tablo 12.** $R^2$ için Kolmogorov-Smirnov testi (Kolmogorov-Smirnov test for $R^2$)

| $R^2$ | Doğrusal Regresyon | Karar ağacı regresyon | GBT regresyon | Rastgele orman regresyon | İzotonik regresyon |
|---|---|---|---|---|---|
| Kolmogorov-Smirnov test değeri | 0,311 | 0,206 | 0,143 | 0,185 | 0,246 |
| p değeri | 0,235 | 0,719 | 0,970 | 0,824 | 0,504 |

**Tablo 13.** Algoritma RMSE sonuçlarının istatistiksel karşılaştırılması (Statistically comparison of the algorithms for RMSE)

| RMSE | Doğrusal regresyon | Karar ağacı regresyon | GBT regresyon | Rastgele orman regresyon | İzotonik regresyon |
|---|---|---|---|---|---|
| Doğrusal regresyon | - | <0,01 | <0,01 | <0,01 | <0,01 |
| Karar ağacı regresyon | <0,01 | - | 0,09 | <0,01 | <0,01 |
| GBT regresyon | <0,01 | 0,09 | - | <0,01 | <0,01 |
| Rastgele orman regresyon | <0,01 | <0,01 | <0,01 | - | <0,01 |
| İzotonik regresyon | <0,01 | <0,01 | <0,01 | <0,01 | - |

**Tablo 14.** Algoritma $R^2$ sonuçlarının istatistiksel karşılaştırılması (Statistically comparison of the algorithms for $R^2$)

| $R^2$ | Doğrusal regresyon | Karar ağacı regresyon | GBT regresyon | Rastgele orman regresyon | İzotonik regresyon |
|---|---|---|---|---|---|
| Doğrusal regresyon | - | <0,01 | <0,01 | <0,01 | <0,01 |
| Karar ağacı regresyon | <0,01 | - | 0,12 | <0,01 | <0,01 |
| GBT regresyon | <0,01 | 0,12 | - | <0,01 | <0,01 |
| Rastgele orman regresyon | <0,01 | <0,01 | <0,01 | - | <0,01 |
| İzotonik regresyon | <0,01 | <0,01 | <0,01 | <0,01 | - |

sonuçlardan daha iyidir. Elde edilen sonuçlar ve sonuçlar arası istatistiksel analizler dikkate alındığında en iyi sonucu veren algoritmanın rastgele orman regresyon algoritması olduğu görülmektedir.

## 6. Sonuçlar (Conclusions)

Bu çalışmada ikinci el araba fiyatları gerçek hayat verisi olarak internet üzerindeki ilanlardan toplanmıştır. Bu çalışma için toplanan veri kümesindeki araç sayısı 120000 civarındadır ve bu büyüklükteki bir veri için geleneksel yöntemler yetersiz ve yavaş kalmaktadır. Bu yüzden büyük veri için uygun bir araç olan Apache Spark kullanılmıştır. Hem fiyat üzerinden hem de arabaların özellikleri üzerinden en çok kullanılan markalar, en çok kullanılan modeller, vites türü, yakıt türü gibi çeşitli analizler yapılmıştır. Bu analizlere ek olarak doğrusal regresyon, karar ağacı regresyonu, GBT regresyon, rastgele orman regresyonu ve izotonik regresyon olmak üzere toplam beş farklı makine öğrenme algoritması kullanılarak araçların fiyat tahminleri gerçekleştirilmiştir. Yapılan analizler sonucunda araç fiyatlarını en yakın olarak tahmin eden algoritma rastgele orman regresyonu algoritması çıkmıştır. Rastgele orman algoritması en düşük RMSE ve en yüksek $R^2$ değerlerine sahiptir. Rastgele orman algoritmasının diğer algoritmalara olan üstünlüğünü istatistiksel olarak test etmek için öncelikle sonuçların normal dağılıma uyup uymadığı Kolmogorov-Smirnov testi ile kontrol edilmiştir. Kolmogorov-Smirnov testi sonucunda normal dağılıma uygun olduğu saptanmıştır. Bu sayede algoritmaların farklılıklarını istatistiksel olarak karşılaştırmak için 2-örnek-t-testi kullanılmıştır. Bu test sonucunda rastgele orman algoritmasının diğer algoritmalardan üstünlüğü istatistiksel olarak anlamlı çıkmıştır. Bu çalışmaya göre rastgele orman algoritması ikinci el araç fiyat tahmini için kullanılabilecek en iyi algoritmadır.

## Kaynaklar (References)


1. Elshawi R., Sakr S., Talia D., Trunfio P., Big Data Systems Meet Machine Learning Challenges: Towards Big Data Science as a Service, Big Data Research, 14, 1–11, 2018.
2. Lu R., Zhu H., Liu X., Liu J.K., Shao J., Toward efficient and privacy-preserving computing in big data era, IEEE Network, 28 (4), 46–50, 2014.
3. García S., Ramírez-Gallego S., Luengo J., Benítez J.M., Herrera F., Big data preprocessing: methods and prospects, Big Data Analytics, 1 (1), 9, 2016.
4. Concolato C.E., Chen L.M., Data Science: A New Paradigm in the Age of Big-Data Science and Analytics, New Mathematics and Natural Computation, 13 (02), 119–143, 2017.
5. Reyes-Ortiz J.L., Oneto L., Anguita D., Big Data Analytics in the Cloud: Spark on Hadoop vs MPI/OpenMP on Beowulf, Procedia Computer Science, 53, 121–130, 2015.
6. Işık K., Ulusoy S.K., Determining the factors that affect the production time in the metal industry utilizing data mining methods, Journal of the Faculty of Engineering and Architecture of Gazi University, 36 (4), 1949–1962, 2021.
7. Apache Spark™ - Lightning-Fast Cluster Computing
8. Duque Barrachina A., O'Driscoll A., A big data methodology for categorising technical support requests using Hadoop and Mahout, Journal of Big Data, 1 (1), 1, 2014.
9. Sarker I.H., Machine Learning: Algorithms, Real-World Applications and Research Directions, SN Computer Science, 2 (3), 160, 2021.
10. Mohammed M., Khan M.B., Bashier E.B.M., Machine Learning: Algorithms and Applications. CRC Press: Boca Raton, 2016.
11. Portugal I., Alencar P., Cowan D., The use of machine learning algorithms in recommender systems: A systematic review, Expert Systems with Applications, 97, 205–227, 2018.
12. Ahmed H., Younis E.M., Ali A.A., Predicting Diabetes using Distributed Machine Learning based on Apache Spark, 2020 International Conference on Innovative Trends in Communication and Computer Engineering (ITCE), 44–49, 2020.
13. Oo M.C.M., Thein T., An efficient predictive analytics system for high dimensional big data, Journal of King Saud University - Computer and Information Sciences, 2019.
14. Río S. del, López V., Benítez J.M., Herrera F., On the use of MapReduce for imbalanced big data using Random Forest, Information Sciences, 285, 112–137, 2014.
15. Sağlamlar H., Multi center polyhedral conic classifiers that can classify complex data, Journal of the Faculty of Engineering and Architecture of Gazi University, 36 (4), 1817–1830, 2021.
16. HimaBindu G., Raghu Kumar Ch., Hemanand Ch., Rama Krishna N., Hybrid clustering algorithm to process big data using firefly optimization mechanism, Materials Today: Proceedings, 2020.

**17.** Tao Q., Gu C., Wang Z., Jiang D., An intelligent clustering algorithm for high-dimensional multiview data in big data applications, Neurocomputing, 393, 234–244, 2020.

**18.** Alnafessah A., Casale G., Artificial neural networks based techniques for anomaly detection in Apache Spark, Cluster Computing, 1–16, 2019.

**19.** Lu W., Improved K-Means Clustering Algorithm for Big Data Mining under Hadoop Parallel Framework, Journal of Grid Computing, 18 (2), 239–250, 2020.

**20.** Cui X., Zhu P., Yang X., Li K., Ji C., Optimized big data K-means clustering using MapReduce, The Journal of Supercomputing, 70 (3), 1249–1259, 2014.

**21.** Shang H., Lu D., Zhou Q., Early warning of enterprise finance risk of big data mining in internet of things based on fuzzy association rules, Neural Computing and Applications, 2020.

**22.** Moens S., Aksehirli E., Goethals B., Frequent Itemset Mining for Big Data, 2013 IEEE International Conference on Big Data, 111–118, 2013.

**23.** Zhang F., Liu M., Gui F., Shen W., Shami A., Ma Y., A distributed frequent itemset mining algorithm using Spark for Big Data analytics, Cluster Computing, 18 (4), 1493–1501, 2015.

**24.** Nodarakis N., Sioutas S., Tsakalidis A.K., Tzimas G., Large Scale Sentiment Analysis on Twitter with Spark., EDBT/ICDT Workshops, 1–8, 2016.

**25.** El Alaoui I., Gahi Y., Messoussi R., Chaabi Y., Todoskoff A., Kobi A., A novel adaptable approach for sentiment analysis on big social data, Journal of Big Data, 5, 12, 2018.

**26.** Hasan R.A., Alhayali R.A.I., Zaki N.D., Ali A.H., An adaptive clustering and classification algorithm for Twitter data streaming in Apache Spark, Telkomnika, 17 (6), 3086–3099, 2019.

**27.** Altintaş V., Albayrak M., Topal K., Topic modeling with latent Dirichlet allocation for cancer disease posts, Journal of the Faculty of Engineering and Architecture of Gazi University, 36 (4), 2183–2196, 2021.

**28.** Syed D., Refaat S.S., Abu-Rub H., Performance evaluation of distributed machine learning for load forecasting in smart grids, 2020 Cybernetics & Informatics (K&I), 1–6, 2020.

**29.** Taşyürek M., Çelik M., FastGTWR: A fast geographically and temporally weighted regression approach, Journal of the Faculty of Engineering and Architecture of Gazi University, 36 (2), 715–726, 2021.

**30.** Arslan S., Aslan S., A new lattice based artificial bee colony algorithm for EEG noise minimization, Journal of the Faculty of Engineering and Architecture of Gazi University, 38 (1), 15–28, 2022.

**31.** Xu Y., Liu H., Long Z., A distributed computing framework for wind speed big data forecasting on Apache Spark, Sustainable Energy Technologies and Assessments, 37, 100582, 2020.

**32.** Manogaran G., Lopez D., Spatial cumulative sum algorithm with big data analytics for climate change detection, Computers & Electrical Engineering, 65, 207–221, 2018.

**33.** Montgomery D.C., Peck E.A., Vining G.G., Introduction to linear regression analysis. John Wiley & Sons, 2012.

**34.** Özel S.Ö., Çabuk S., Estimation of ill-posed linear deterministic regression model: generalized maximum entropy and bayesian approach, Journal of the Faculty of Engineering and Architecture of Gazi University, 37 (2), 815–824, 2022.

**35.** Bisong E., Linear Regression, in Building Machine Learning and Deep Learning Models on Google Cloud Platform: A Comprehensive Guide for Beginners, Bisong E, Editor. Apress: Berkeley, CA. 231–241, 2019.

**36.** Xu M., Watanachaturaporn P., Varshney P.K., Arora M.K., Decision tree regression for soft classification of remote sensing data, Remote Sensing of Environment, 97 (3), 322–336, 2005.

**37.** Gökdemr A., Çalhan A., Deep learning and machine learning based anomaly detection in internet of things environments, Journal of the Faculty of Engineering and Architecture of Gazi University, 37 (4), 1945–1956, 2022.

**38.** Veri Madenciliği'nde Karar Ağaçları, MSHOWTO Topluluğu ve Bilişim Portalı, 2020.

**39.** Zhang Y., Haghani A., A gradient boosting method to improve travel time prediction, Transportation Research Part C: Emerging Technologies, 58, 308–324, 2015.

**40.** Shoaran M., Haghi B.A., Taghavi M., Farivar M., Emami-Neyestanak A., Energy-efficient classification for resource-constrained biomedical applications, IEEE Journal on Emerging and Selected Topics in Circuits and Systems, 8 (4), 693–707, 2018.

**41.** Segal M.R., Machine Learning Benchmarks and Random Forest Regression, 2004.

**42.** Barlow R.E., Brunk H.D., The isotonic regression problem and its dual, Journal of the American Statistical Association, 67 (337), 140–147, 1972.

**43.** Isotonic regression, Wikipedia, 2020.