T.C.
SAKARYA ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

WEB KAZIMA VE MAKİNE ÖĞRENMESİ YÖNTEMLERİ KULLANILARAK FİYAT TAHMİNLEME: İKİNCİ EL ARAÇ PİYASASINDA BİR ÖRNEK

YÜKSEK LİSANS TEZİ

Seda YILMAZ

Bilişim Sistemleri Mühendisliği Anabilim Dalı

AĞUSTOS 2023

T.C.
SAKARYA ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

WEB KAZIMA VE MAKİNE ÖĞRENMESİ YÖNTEMLERİ KULLANILARAK FİYAT TAHMİNLEME: İKİNCİ EL ARAÇ PİYASASINDA BİR ÖRNEK

YÜKSEK LİSANS TEZİ

Seda YILMAZ

Bilişim Sistemleri Mühendisliği Anabilim Dalı

Tez Danışmanı: Doç. Dr. İhsan Hakan SELVİ

AĞUSTOS 2023

Seda YILMAZ tarafından hazırlanan "WEB KAZIMA VE MAKİNE ÖĞRENMESİ YÖNTEMLERİ KULLANILARAK FİYAT TAHMİNLEME: İKİNCİ EL ARAÇ PİYASASINDA BİR ÖRNEK" adlı tez çalışması 09.08.2023 tarihinde aşağıdaki jüri tarafından oy birliği/oy çokluğu ile Sakarya Üniversitesi Fen Bilimleri Enstitüsü Bilişim Sistemleri Mühendisliği Anabilim Dalı Yüksek Lisans tezi olarak kabul edilmiştir.

Tez Jürisi

Jüri Başkanı :	**Prof. Dr. Orhan ER** Bakırçay Üniversitesi	.....
Jüri Üyesi :	**Doç. Dr. İhsan Hakan SELVİ** Sakarya Üniversitesi	.....
Jüri Üyesi :	**Prof. Dr. Numan ÇELEBİ** Sakarya Üniversitesi	.....

ETİK İLKE VE KURALLARA UYGUNLUK BEYANNAMESİ

Sakarya Üniversitesi Fen Bilimleri Enstitüsü Lisansüstü Eğitim-Öğretim Yönetmeliğine ve Yükseköğretim Kurumları Bilimsel Araştırma ve Yayın Etiği Yönergesine uygun olarak hazırlamış olduğum "WEB KAZIMA VE MAKİNE ÖĞRENMESİ YÖNTEMLERİ KULLANILARAK FİYAT TAHMİNLEME: İKİNCİ EL ARAÇ PİYASASINDA BİR ÖRNEK" başlıklı tezin bana ait, özgün bir çalışma olduğunu; çalışmamın tüm aşamalarında yukarıda belirtilen yönetmelik ve yönergeye uygun davrandığımı, tezin içerdiği yenilik ve sonuçları başka bir yerden almadığımı, tezde kullandığım eserleri usulüne göre kaynak olarak gösterdiğimi, bu tezi başka bir bilim kuruluna akademik amaç ve unvan almak amacıyla vermediğimi ve 20.04.2016 tarihli Resmi Gazete'de yayımlanan Lisansüstü Eğitim ve Öğretim Yönetmeliğinin 9/2 ve 22/2 maddeleri gereğince Sakarya Üniversitesi'nin abonesi olduğu intihal yazılım programı kullanılarak Enstitü tarafından belirlenmiş ölçütlere uygun rapor alındığını, çalışmamla ilgili yaptığım bu beyana aykırı bir durumun ortaya çıkması halinde doğabilecek her türlü hukuki sorumluluğu kabul ettiğimi beyan ederim.

30/05/2023

Seda YILMAZ

Aileme...

TEŞEKKÜR

Bu çalışmanın gerçekleştirilmesinde değerli bilgi ve deneyimlerini benimle payşalaşan ve her türlü sorumda yardımcı olan, desteklerini esirgemeyen danışman hocam Doç. Dr. İhsan Hakan SELVİ'ye sonsuz teşekkür ve saygılarımı sunarım.

Çalışmalarım ve eğitim hayatım boyunca destekleriyle beni hiçbir zaman yalnız bırakmayan başta canım annem olmak üzere aileme, Cemal dayıma ve Yasemin teyzeme de sonsuz teşekkürler ederim.

Seda Yılmaz

İÇİNDEKİLER

Sayfa

KISALTMALAR

ABR **:** Adaptive Boosting Regression
API **:** Application Programming Interface
CART **:** Classification and Regression Trees
GBR **:** Gradient Boosting Regression
IDE **:** Integrated Development Environment
KNN **:** K-Nearest Neighbors
MSE : Mean Squared Error
PCA : Principal Component Analysis
RFR : Random Forest Regression
RMSE : Root Mean Squared Error
SVR : Support Vector Regression
URL : Uniform Resource Locator

SİMGELER

$\mathbf{R^2}$: Determinasyon katsayısı

TABLO LİSTESİ

Sayfa

ŞEKİL LİSTESİ

Sayfa

WEB KAZIMA VE MAKİNE ÖĞRENMESİ YÖNTEMLERİ KULLANILARAK FİYAT TAHMİNLEME: İKİNCİ EL ARAÇ PİYASASINDA BİR ÖRNEK

ÖZET

Teknolojinin gelişmesi ile birlikte günümüzde veri miktarı ve trafiği önemli ölçüde arttı. Dolayısıyla veriyi toplamak ve anlamlandırabilmek de oldukça önemli hale geldi. Ancak ihtiyaç duyulan veriler her zaman bir kaynakta toplanıp sunulmuyor olabilir. Bu nedenle web kazıma yöntemleri kullanılarak, ihtiyaç duyulan veriler ilgili kaynaklardan toplanabilmektedir.

Bu çalışmada web kazıma teknikleri kullanılarak toplanan ikinci el araç satış verilerinin makine öğrenmesi algoritmaları kullanılarak analiz edilmesi ve fiyat tahminleme modeli oluşturulması hedeflenmiştir. Analiz için ihtiyaç duyulan veriler belirlenip Selenium ve BeautifulSoup kullanılarak toplanmıştır. Çalışmada kullanılmak üzere aynı özelliklere sahip, farklı sayıda veri içeren iki adet veri seti elde edilmiştir. Her iki veri seti de biri hedef değişken olmak üzere 25 özellikten oluşmaktadır. Veri seti 1, 5557 satır veri içerirken veri seti 2 ise 11688 satır veri içermektedir. İki veri seti de çeşitli veri ön işleme aşamalarından geçirilerek analize hazırlanmıştır. Toplanan veriler özellik seçimi ve boyut indirgeme için Lasso regresyon ve PCA analizi, hiperparametre ayarlaması yapmak için GridSearchCV yöntemi uygulanarak makine öğrenmesi algoritmaları ile değerlendirilmiştir. Lasso regresyon sonrası veri seti 1 için özellik sayısı 11'e indirgenirken veri seti 2 için 9'a indirgenmiştir. PCA analizi sonrası ise her iki veri seti için de özellik sayısı 7'ye indirgenmiştir.

Analizde Rastgele Orman Regresyon, K-en Yakın Komşu Regresyon, Gradyan Artırma Regresyon, AdaBoost Regresyon, Destek Vektör Regresyon ve XGBoost Regresyon algoritmaları kullanılmıştır. Elde edilen analiz sonuçları Ortalama Kare Hata (MSE), Kök Ortalama Kare Hata (RMSE) ve Determinasyon Katsayısı (R^2) ile birlikte değerlendirilmiştir.

Veri seti 1 için sonuçlar incelendiğinde en iyi sonucu veren model 0.973 R^2, 0.026 MSE ve 0.161 RMSE değerleri ile XGBoost Regresyon olmuştur. Bu modeli Gradyan Artırma Regresyon modeli takip etmektedir. Veri seti 2 için sonuçlar incelendiğinde ise en iyi sonucu veren model 0.978 R^2, 0.021 MSE ve 0.145 RMSE değerleri ile K-en Yakın Komşu Regresyon olmuştur. Bu modeli XGBoost Regresyon modeli takip etmektedir. Analiz sonucunda her iki veri seti için de en kötü sonuçları veren algoritmanın AdaBoost Regresyon olduğu görülmüştür.

PRICE PREDICTION USING WEB SCRAPING AND MACHINE LEARNING METHODS: AN EXAMPLE IN THE USED CAR MARKET

SUMMARY

With the developing technology, the size of the data and the data traffic are increasing day by day. Therefore, data analysis is of great importance in many different sectors and functions today. When data analysis is done correctly, it provides significant benefits in almost every field. It enables businesses to achieve better results in decision-making. Accordingly, it provides important benefits such as competitive advantage, customer satisfaction, operational efficiency, and better management of risks. Data collection and interpretation are important. However, data that is not collected and interpreted correctly only causes confusion. Therefore, data collection is as important as analysis. The required data may not always be provided directly. Therefore, by using web scraping methods, the needed data can be collected from relevant sources.

Websites contain large amounts of data. Instead of collecting this data manually, web scraping makes it possible to collect large amounts of data automatically. With web scraping, the collection of large amounts of data is automated and the data is made more useful to the user.

Identifying needs is important in web scraping. First of all, it should be determined which data will be collected from which website. Different tools and libraries are available for web scraping. Libraries like BeautifulSoup and Scrapy in Python are popular and frequently used methods for web scraping. By doing research, the most suitable method for web scraping can be determined.

After data is collected, some preprocessing steps may be required before starting the analysis. By examining the data set, missing or incorrect data is corrected or removed. Scaling is done for data of different scales and a clean data set is obtained by removing redundant or repetitive data. Appropriate methods should be determined for the analysis of the obtained data set. These methods vary depending on the characteristics of the dataset and the intended output. Machine learning, statistical analysis, data mining or other analytical methods can be used.

In this study, it is aimed to analyze the used car sales data collected using web scraping techniques by using machine learning algorithms and to create a price prediction model. The data used in the study were collected with codes written in the Python programming language from a used car sales website and saved in a database. Selenium and BeautifulSoup were used during web scraping for data collection.

In the web scraping process, the URLs of the requested products were first retrieved from the target website via Selenium. The source codes of the pages whose URLs were taken were collected and the desired parts were parsed using BeautifulSoup. The data collected with Numpy and Pandas were manipulated, converted into appropriate

format and tabulated. The collected data was written to the database. Used Pymysql and Sqlalchmy for database connection.

Data collected; price, brand, model, series, year, km, type, information from, fuel, gear, engine power, engine capacity. In addition, the following information was collected from the template showing the damage status; right rear fender, rear hood, left rear fender, right rear door, right front door, roof, left rear door, left front door, right front fender, engine hood, left front fender, front bumper and rear bumper. While these fields are being filled, 0 values are assigned if the relevant part of the vehicle is original, 1 if it is painted, 2 if it is changed, and 3 if it is not specified.

Within the scope of the study, two data sets with different data numbers were obtained in order to make comparisons during the analysis. Both datasets consist of 25 features. While collecting the data, the dates between 10.02.2023 and 31.03.2023 were taken as a basis. Dataset 1 contains 5557 rows of data, while Dataset 2 contains 11688 rows of data.

The collected data were analyzed and initially, some fields that were null in the columns reflecting the damage status were filled with the default (unspecified=3) value. Then, categorical data were converted into numerical data using the encoder method to be ready for analysis. The price variable was set as the target variable.

Lasso regression was used to determine which of the remaining 24 features would be used in the analysis. Lasso regression is a regression technique used in feature selection and model parameter estimation. Unlike traditional regression analysis, it allows the coefficient of unimportant variables to be reduced to zero. With this feature, it can reduce overfitting problems and make the model simpler and more interpretable. After applying Lasso, 11 features were selected in Dataset 1 result, and 9 features were selected in Dataset 2 result. Afterward, PCA analysis for size reduction was applied and the variables to be included in the analysis were determined. Principal Component Analysis (PCA) is a statistical technique used for dimension reduction in multivariate datasets. PCA is used to analyze the relationships between the variables in the data set and to create new variables that best explain these relationships. As a result of the PCA analysis, the number of features was reduced to 7 for both data sets.

Training and test data were separated for use in machine learning algorithms. K-Fold cross-validation was used for this process. In the K-Fold cross-validation method, the dataset is divided into k different subsets. It is ensured that each piece is used as both training and test data. In this way, the errors that may occur due to its distribution are minimized. In this study, analyses were made by choosing k value of 10.

Six different machine-learning algorithms were used in the study. Both the data set with the lower number of data and the data set with the higher number of data were run with each of these algorithms. GridSearchCV was used to determine the best parameters in the algorithms.

Random Forest Regression, K-Nearest Neighbor Regression, Gradient Boosting Regression, AdaBoost Regression, Support Vector Regression, and XGBoost Regression algorithms were used in the analysis. The obtained analysis results were evaluated together with Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Coefficient of Determination (r-squared), and learning curve graphs.

When the results for data set 1 were examined, the model that gave the best results was XGBoost Regression with 0.973 R^2, 0.026 MSE, and 0.161 RMSE values. This model is followed by the Gradient Boosting Regression model. When the results for data set 2 were examined, the model that gave the best results was K-Nearest Neighbor Regression with 0.978 R^2, 0.021 MSE, and 0.145 RMSE values. This model is followed by the XGBoost Regression model. As a result of the analysis, it was seen that the algorithm that gave the worst results for both data sets was AdaBoost Regression.

1. GİRİŞ

Gelişen teknoloji ile birlikte veri trafiği ve verinin boyutu günden güne artmaktadır. Dolayısıyla günümüzde veri analizi, birçok farklı sektörde ve işlevde büyük önem taşımaktadır. Veri analizi doğru şekilde yapıldığında hemen hemen her alanda önemli ölçüde etkilidir. İşletmelerin karar almada daha iyi sonuçlar elde etmelerini ve buna bağlı olarak rekabet avantajı, müşteri memnuniyeti, operasyonel verimlilik ve riskleri daha iyi yönetme gibi önemli faydalar sağlamaktadır.

Ancak doğru şekilde toplanmamış ve anlamlandırılmamış veri yalnızca karmaşaya neden olmaktadır. Bu nedenle verinin analizi kadar toplanması da önem taşımaktadır. İhtiyaç duyulan veriler her zaman bir kaynakta toplanıp sunulmuyor olabilir. Bu nedenle web kazıma yöntemleri kullanılarak, ihtiyaç duyulan veriler ilgili kaynaklardan toplanabilmektedir.

Web siteleri büyük miktarda veri barındırmaktadır. Bu verileri elle çıkarma işlemi yerine web kazıma çok miktarda veriyi otomatik olarak çıkarmayı mümkün hale getirir. Web kazıma ile hem çok miktarda verinin toplanması otomatikleştirilmiş olur hem de veriler kullanıcı için daha kullanışlı bir biçime getirilir. Ancak çoğu durumda web kazıma basit bir işlem değildir. Web siteleri farklı şekil ve biçimler barındırır. Bu nedenle web kazıyıcılar işlevsellik ve özelliklere göre değişmektedir.

Web kazımada ihtiyaçların belirlenmesi önemlidir. İlk olarak hangi web sitesinden hangi verilerin çekileceği belirlenmelidir. Web kazıma işlemi için farklı araç ve kütüphaneler mevcuttur. Python dilinde BeautifulSoup ve Scrapy gibi kütüphaneler web kazıma için popüler ve sık kullanılan yöntemlerdendir. Araştırma yapılarak en uygun seçenek belirlenebilir.

Veriler toplandıktan sonra analizlere başlanmadan önce bazı ön işleme adımları uygulamak gerekebilir. Veri seti incelenerek eksik veya hatalı veriler düzeltilir ya da çıkartılır. Farklı ölçeklerdeki verilerin aynı skalaya getirilmesi için ölçeklendirme yapılır ve gereksiz veya tekrarlayan veriler kaldırılarak temiz bir veri seti elde edilir.

Elde edilen veri setinin analizi için uygun yöntemler belirlenmelidir. Veri setinin özelliklerine ve hedeflenen çıktıya bağlı olarak bu yöntemler değişmektedir. Makine öğrenmesi, istatistiksel analiz, veri madenciliği veya diğer analitik yöntemleri kullanılabilir.

Bu çalışmada ikinci el araç piyasasında araçların özelliklerine göre fiyatlarının tahmin edilmesi hedeflenmiştir. Bu doğrultuda elde edilmek istenen veriler belirlenmiş ve web kazıma yöntemleri ile bu veriler toplanmıştır. Toplanan veriler, veri ön işleme adımlarından geçirilmiş ve makine öğrenmesi algoritmaları ile analiz edilmiştir.

2. LİTERATÜR ARAŞTIRMASI

Çok sayıda web sitesi büyük miktarlarda veri içerir. Bu bilgileri istatistiksel özetlerle elde etmek şahıslar ve firmalar için önem taşır (Milev, 2017). Büyük veri ve bu verilerin analizi olmasaydı web kazımanın gelişmeyeceği hatta hiç ihtiyaç duyulmayacağı bile varsayılabilir. Bu nedenle, veri bilimi web kazımada kilit bir rol oynar. Herhangi bir bilgiye ulaşmak amacıyla web sitelerini ziyaret ederken, herkes büyük miktarda veri ile dolu web sayfalarıyla karşı karşıya kalır. İnternet, farklı alanlar için muazzam miktarda veriyle doludur (Broucke ve Baesens, 2018). Çoğu zaman veriler doğrudan kopyalanamaz. Bu nedenle, verileri analiz etmek için mümkün olduğunca az zaman ve çaba harcayacak becerilere ihtiyaç vardır. Web sayfası sahipleri kullanıcıların ilgisini çekmek için web sitelerini bolca resim ve animasyonla zenginleştirmeye çalışırlar. Kasıtlı veya değil, tüm bunlar web sayfalarının doğrudan veri indirmesini de engeller. Bilgiler, indirmesi pek kolay olmayacak şekilde yerleştirilmiştir ve indirmesi kolay olsa bile işlenmesi zor olacaktır. Web kazımanın veri bilimine girdiği yer burasıdır (Khder, 2021).

Web kazıma yöntemi yalnızca bilgi toplamaya yardımcı olmakla kalmaz, aynı zamanda daha sonraki analizlerde kolaylık sağlamak için bilgiyi yapılandırır. Elbette web sayfalarından manuel olarak da veri toplanabilir. Ancak bu işlem çok fazla zaman alır ve karmaşık site yapıları, bilgilerin sitelerden bu kadar kolay kopyalanmasına izin vermeyebilir. Bu sebeplerden dolayı web kazıma yöntemi ortaya çıkmıştır (Banerjee, 2014).

Web kazıma gibi yenilikler sayesinde, girişimciler yeni iş projeleri açmak, mevcut işletmeleri genişletmek ve geliştirmek ve ürün ve hizmetlerin performansını, dağıtımını, kârlılığını, rekabet gücünü ve uygunluğunu geliştirmek için daha fazla fırsata sahiptir. Örneğin, web kazıma yoluyla elde edilebilecek büyük miktarda veri yardımıyla istatistiksel analizler yapabilir ve iş yaparken hangi yönde ilerlemeniz gerektiğini anlayabilirsiniz. Ya da aradığınız özelliklerdeki ürün için ortalama fiyat aralığını bulabilirsiniz. Web kazıma, bir müşteri tabanı oluşturmaya yardımcı olur. Web kazıma yardımıyla, bir web sayfasını veya sosyal ağı ziyaret etmiş kişileri

bulabilir ve bu aynı kişilere girişimcinin temsil ettiği ürünün aynısını sunabilirsiniz. Ayrıca diğer firmaların fiyatlarını, çeşitlerini, promosyon ve indirimlerini, yeni gelenleri ve yenilikleri ve bir işletmeyi daha rekabetçi hale getirebilecek her şeyi takip edebilirsiniz. Müşteri incelemeleri, yorumları, derecelendirmeleri veya değerlendirmeleri, iş geliştirmede önemli bir rol oynar. Birçok kişi satın almadan önce başkalarının deneyimlerine güvenir ve satın almadan önce bir hizmet veya ürün hakkında canlı bir inceleme okur (Boegershausen ve ark., 2020). Web kazıma ayrıca bu tür verileri toplayabilir, anahtar kelimelere göre yapılandırabilir ve bitmiş bir sonuç sağlayabilir. Bu şekilde web kazıma, pazarlama yollarının uyarlanmasına ve gerçek, alakalı ve sağlam analize dayalı reklamların geliştirilmesine yardımcı olur.

Web kazıma biliyor olmak büyük verilerle çalışırken analiz için birçok fırsat sağlar. Artık pek çok web sitesi kullanıcılara indirilebilen ve kullanılabilen yapılandırılmış veri depolarına erişim sağlayan bir API (Application Programming Interface) sağlamaktadır. Eğer web sitesi API'ye erişim izni sağlıyorsa kazımanın bir anlamı yoktur. Bu nedenle web kazıma işlemi yapmadan önce API'nin genel erişime açık olup olmadığını kontrol edilmelidir. Web kazımanın şu anki en önemli kullanımlarından biri, işletmelerin rakiplerinin fiyatlandırma faaliyetlerini izlemesidir. Fiyatlandırma, nispeten kısa zaman ölçeklerinde ve minimum manuel çabayla tüm bir site genelinde oluşturulabilmektedir (Haddaway, 2015). Eğer web sayfası API'ye erişim sağlamıyorsa, verileri elde etmenin başka yolu yoktur. Burada en iyi çözüm web kazımadır. Bunların dışında web sitelerinin bir API sağladığı ancak bu hizmetin ücretli olduğu durumlar olabilir. Bu da birçok kullanıcıyı web kazıma yöntemine başvurmaya zorlar. Ya da sağlanan API hız olarak sınırlandırılabilir ve kullanıcılar günde bir veya saatte bir verileri belirli bir sıklıkta kullanabilir. Web kazıma yöntemini kullanmanın bir başka nedeni de web sayfasının API'den çok daha fazla bilgi sağlaması ve çoğu zaman birçok kullanıcının istediği bilgilerin kapatılması olabilir. Burada temel fikir web sayfalarındaki verilerin bilinen yollardan biriyle bulup indirmenin yanı sıra bunları kullanmanın ve onlarla çalışmanın mümkün olduğudur (Henrys, 2021).

Doğru araba fiyatı tahmini, uzman bilgisi gerektirir, çünkü fiyat genellikle birçok ayırt edici özelliğe ve faktöre bağlıdır. Bir otomobilin fiyatını doğru bir şekilde tahmin edebilmek için otomobilin beygir gücü, motor kapasitesi veya şanzıman gibi teknik

özelliklerinden farklı olması gereken birçok boyutun yanı sıra, bu alanda uzman bazı kişilerin bilgisinin gerekli olduğu gösterilmiştir (Gegic ve ark., 2019).

Pandey ve ark. (2020), yaptıkları çalışmada makine öğrenmesi algoritmalarının bu tür alanlarda karşılaşılan sorunları çözmek için kullanılabilecek iyi çözümler olduğunu söylemektedir. Pandey ve arkadaşlarına göre makine öğrenmesi algoritmaları ve teknikleri, ikinci el araç fiyat tahmini konusu gibi bu tür sorunlara iyi bir çözüm sunabilir.

İkinci el araç fiyat tahmini konsepti büyük bir potansiyel göstermekte ve katlanarak büyümektedir. Doğru bir ikinci el fiyat değerlendirmesi, ikinci el fiyat kavramı pazarlama alanının sağlıklı bir şekilde gelişmesi için çok önemlidir (Chen ve ark., 2020). Chen ve arkadaşları çalışmalarında büyük veri ve kullanım alanlarına dikkat çekerek kapsamlı bir çalışma ortaya koymuşlardır. Kullanılan verinin boyutu arttıkça taşıdığı önemin de arttığını yazılarında vurgulamaktadırlar. Çoklu doğrusal regresyon algoritmasını uygularken üç ana şeyden etkilendiğini fark ederler. Bunlardan birincisi algoritmanın kendisi, ikincisi açıklayıcı değişkenlerin sayısı ve son olarak toplam örnek sayısıdır. Çalışmaları, çoklu doğrusal regresyon ve rastgele orman tekniklerinin modelleme bölümünde farklı karmaşıklıklarla nasıl çalıştığını göstermektedir. Sonuç olarak çalışmada ikinci el araç fiyat tahmini ile uğraşırken tek amacın en uygun modeli gerçekleştirmek olduğu savunulmaktadır (Chen ve ark., 2020).

Asghar ve ark. (2021), pazarlamada kullanılan araçların değerine göre satış fiyatlarının tahmin edilmesi için bazı farklı yolların kullanılmasını önermektedir. Önerdikleri uygulamalar, hem alıcıya hem de satıcıya ikinci el araç alım satımı konusunda destek olmakta, kendileri için en uygun fiyatları tahmin edebilmekte ve hem kişisel hem de ticari açıdan iyi bir tespit yapabilmektedir. Önerilen modelleme performansları, çalışmalarının etkili ve verimli yöntemler ve stratejiler içerdiğini yansıtmaktadır. Önerdikleri çalışmalarda bir makine öğrenmesi algoritması olan MLR algoritması üstün bir performans göstermeye yardımcı olmaktadır. Tüm bu bilgiler ışığında, p-değeri yöntemini kullanarak modelleme başarısını ve performanslarını kontrol etmek için bir takım testler uygularlar. Sonuç olarak yaptıkları analizlere göre tahmin puanları verimli ve etkili bir çalışma sergilediklerini göstermektedir (Asghar ve ark., 2021).

3. MATERYAL VE YÖNTEM

Bu çalışmada web kazıma teknikleri kullanılarak web sitesinden verilerin toplanması ve toplanan verilerin makine öğrenmesi algoritmaları ile analiz edilerek fiyat tahminleme yapılması hedeflenmiştir.

3.1. Materyal

Çalışmada kullanılan veriler, ikinci el araç satışı yapılan bir web sitesi üzerinden Python programlama dili ile yazılan kodlar yardımıyla toplanmış ve bir veritabanına kaydedilmiştir.

3.1.1. Web kazımada Python kullanımı

Python veri alanında kullanılan en yaygın programlama dillerinden biridir. Anlaşılması kolay olmasından dolayı tercih edilme oranı yüksektir.

Bu programlama dili genel amaçlıdır, yani çok çeşitli problemler ve görevler için kullanılır. Python çok sayıda avantaja sahiptir ve bu nedenle çeşitli programlama seviyelerindeki geliştiriciler arasında popüler hale gelmiştir. Mantıksal bir sözdizimine sahiptir, yani yapısı ve dolayısıyla kaynak kodunun okunması ve anlaşılması kolaydır. Ayrıca, bu programlama dili esnek ve ölçeklenebilirdir. Bu, gerekirse görevleri genişletmek veya karmaşıklaştırmak için Python dilinin geliştiricilerin çok düzeyli bir kod yapısı oluşturmasına izin verdiği anlamına gelir. Kod yazarken Python düz bir metin dosyası olduğundan, birçok platform bunu yazmak için uygundur. Ayrıca popülaritesi nedeniyle çok sayıda çevrimiçi topluluk vardır. Ayrıca dünyanın her yerinden geliştiriciler Python programlama dilini kullandığı için çeşitli karmaşıklıktaki kodları yazarken başkalarına danışabilir ve yardımcı olabilir.

Bu programlama dilini kullanmanın birçok avantajı nedeniyle, web kazıma için kodlar ve programlar yazma yolunu da bulmuştur. Web kazıma, geliştiriciler için nispeten modern bir eğilimdir, ancak Python programlama dili, web kazıma görevlerini gerçekleştirmek için çok uygundur. Kod yazmaya başlamadan önce, verilerin toplanacağı web sitesini incelemek önemlidir. Python, web kazıma kullanarak veri

toplamak için algoritmalar geliştirmek için çeşitli araçlardan oluşan büyük bir kitaplığa sahiptir.

3.1.2. PyCharm IDE

PyCharm, etkili Python programlaması için gerekli araç setine sahip Python için kullanılan bir IDE'dir. IDE (Tümleşik geliştirme ortamı) kullanıcıların daha hızlı ve daha verimli kod yazmasına yardımcı olan bir ortamdır. Metin düzenlemeye, derlemeye veya yorumlamaya, hata ayıklamaya ve daha fazlasına olanak sağlar. IDE, geliştirme hızının artırılmasına olanak tanır.

3.1.3. Jupyter Notebook

Bir web tarayıcısı üzerinde not defteri formatında kodların çalıştırılabildiği, notlar alınabilen ve istenilen düzenlemelerin gerçekleştirilebildiği bir sunucu-istemci uygulamasıdır. Python dışında başka dilleri de (R, Julia, Ruby vb.) desteklemektedir. Blok mantığında çalışması sayesinde kodun istenilen kısımlarında değişiklikler yapılarak çıktıları gözlemleme imkanı sağlaması sebebiyle çok fazla tercih edilmektedir.

3.1.4. Kullanılan kütüphaneler

Bu çalışmada verilerin toplanması için web kazıma aşamasında Selenium ve Beautifulsoup kullanıldı. Selenium kütüphanesi genel olarak web uygulamalarını test amaçlı otomatikleştirmek için kullanılmaktadır. Ancak yalnızca bunun için değil, tarayıcı üzerinden veri toplamaya da olanak tanımaktadır.

BeautifulSoup kütüphanesi ise HTML dosyalarını işlemek için oluşturulmuş bir kütüphanedir. İlgili kaynak içerisinde yer alan HTML kodlarını ayrıştırarak istenilen kısımların elde edilmesini sağlar.

Verilerin ön işleme ve analiz aşamalarında temel olarak Numpy, Pandas, Scikit-learn ve Mathplotlib kütüphanelerinden yararlanıldı. Numpy, diziler, cebir ve matrisler üzerinde çalışmayı kolaylaştırması ve matematiksel işlemleri gerçekleştirmedeki yeteneği sebebiyle başta veri bilimi olmak üzere istatistik ve matematik alanlarında da sıkça tercih edilen bir kütüphanedir.

Pandas, veri işleme ve analizi için oluşturulan bir Python kütüphanesidir. Bir kaynak üzerindeki verilerin okunması, işlenmesi, filtrelenmesi ve değişikliklerin yapılması

gibi amaçlar için kullanılır. Özellikle Numpy kütüphanesi ile birlikte veri bilimi alanında çok yaygın olarak kullanılmaktadır.

Scikit-learn kütüphanesi Numpy, SciPy Matplotlib kütüphaneleri üzerine kurulan içerisinde birçok istatistiksel model yöntemini hazır bulunduran (sınıflandırma, kümeleme, regresyon algoritmaları) bir Python kütüphanesidir. Yine diğer kütüphaneler ile birlikte veri bilimi alanında kullanımı yaygındır.

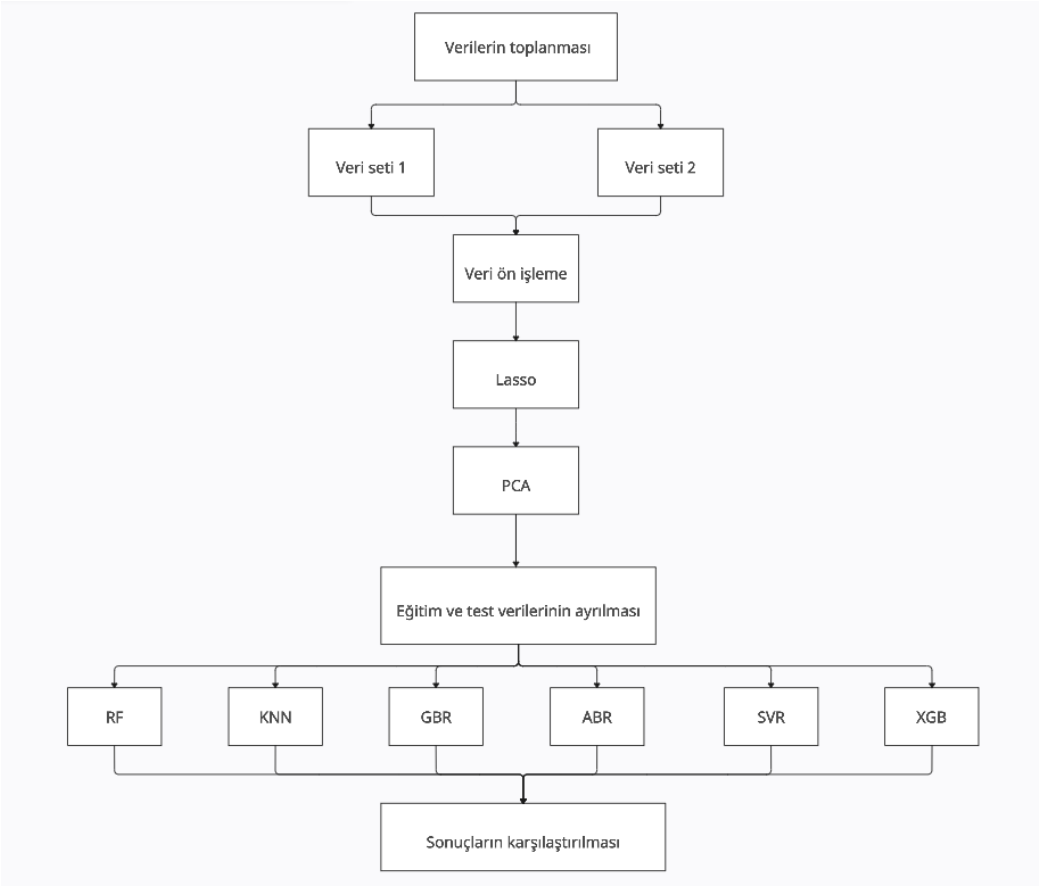
Matplotlib, Python programlama dilinde kullanılan en temel ve yaygın görselleştirme kütüphanesidir. Grafikleri hızlıca sık tercih edilen formatlara da dönüştürerek kaydedebiliyor olması belirgin avantajlarındandır. Bu kütüphane ile veriyi ifade edecek pasta, çubuk, histogram vb. grafiklerin oluşturabilmesi sağlanmaktadır.

3.1.5. Veritabanı

Web kazıma yöntemi ile elde edilen veriler uygun formata getirildikten sonra tablolar halinde MySQL veritabanı üzerine yazıldı. MySQL ücretsiz bir veritabanı sistemidir. Bu işlemler sırasında pymysql ve sqlalchemy kütüphanelerinden faydalanıldı. Bu kütüphaneler sayesinde veritabanındaki ilgili tabloya Python ile erişilebilmekte ve çeşitli select, insert, update, delete işlemleri gerçekleştirilebilmektedir.

3.2. Yöntem

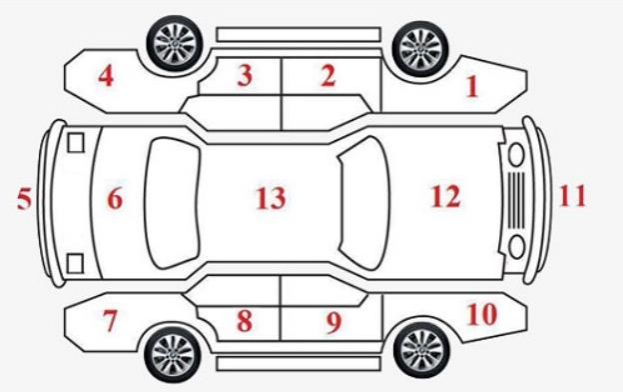
Çalışma kapsamında ikinci el bir araç sitesinden belirlenen kriterler doğrultusunda veriler toplandı. Aşağıdaki adımlar takip edilerek çalışma tamamlandı.



Şekil 3.1. Tahmin metodu

3.2.1. Verilerin toplanması

Öncelikle ikinci el araç satışı yapılan bir web sitesinden araç fiyat tahminlemesi için takip eden verilerin toplanması planlandı; fiyat, marka, model, seri, yıl, km, tip, kimden bilgisi, yakıt, vites, motor gücü, motor hacmi. Bunlara ek olarak ikinci el araç piyasasında önemli bir kriter olan araç hasar durumu da analize katılmak istendi ve sitede araç hasar durumunu gösteren şablon (görsel temsilidir) üzerinden de hasar bilgileri alındı.



Şekil 3.2. Araç şablonu

Bu bilgiler verisetine takip eden sıralama ile eklendi; sağ arka çamurluk, arka kaput, sol arka çamurluk, sağ arka kapı, sağ ön kapı, tavan, sol arka kapı, sol ön kapı, sağ ön çamurluk, motor kaputu, sol ön çamurluk, ön tampon, arka tampon. Bu alanlar doldurulurken aracın ilgili kısmı orjinal ise 0, boyalı ise 1, değişen ise 2, belirtilmemiş ise 3 değerleri atandı.

Web kazıma sürecinde ilk olarak Selenium aracılığıyla hedef web sitesinden istenilen ürünlerin URL'leri alındı. URL'leri alınan sayfaların kaynak kodları toplandı ve BeautifulSoup kullanılarak istenilen kısımlar ayrıştırıldı. Numpy ve Pandas ile çekilen veriler manipüle edilerek uygun formata dönüştürüldü ve tablo haline getirildi. Toplanan veriler veritabanına yazıldı. Veritabanı bağlantısı için pymysql ve sqlalchmy kullanıldı.

3.2.1.1. Veri seti

Çalışma kapsamında analiz esnasında karşılaştırma yapabilmek amacıyla veri sayıları birbirinden farklı olan iki adet veri seti elde edildi.

Her iki veri seti de 25 özellikten oluşmaktadır. Veriler toplanırken 10.02.2023-31.03.2023 tarihleri arası baz alındı. Veri seti 1, 5557 satır veri içerirken veri seti 2 ise 11688 satır veri içermektedir.

3.2.2. Veri ön işleme

Toplanan veriler incelendi ve ilk olarak hasar durumunu yansıtan kolonlarda boş gelen bazı alanlar varsayılan (belirtilmemiş=3) değerleri ile dolduruldu. Daha sonra analize hazır olabilmesi için kategorik veriler encoder yöntemi kullanılarak sayısal verilere dönüştürüldü.

Tahmin edilmesi hedeflenen fiyat değişkeni hedef değişken olarak atandı. Kalan 24 özellikten hangilerinin analizde kullanılacağının belirlenmesi için Lasso regresyon kullanıldı. Lasso sayesinde önemsiz değişkenlerin katsayısı sıfıra indirildi.

Lasso uygulandıktan sonra Veri seti 1 sonucunda 11 özellik seçilirken veri seti 2 sonucunda ise 9 özellik seçildi.

Lasso regresyon, özellik seçimi ve model parametre tahmininde kullanılan bir regresyon tekniğidir. Geleneksel regresyon analizinden farklı olarak, model parametrelerinin sıfıra yaklaşmasını teşvik eder. Bu özelliği sayesinde, aşırı uyum problemlerini azaltabilir, modeli daha basit ve yorumlanabilir hale getirebilir.

Daha sonra boyut indirgeme amaçlı PCA analizi uygulandı ve analizde yer alacak değişkenler belirlendi. PCA analizi sonucunda her iki veri seti için de özellik sayısı 7'ye indirgendi.

Principal Component Analysis (PCA), çok değişkenli veri setlerinde boyut azaltma için kullanılan istatistiksel bir tekniktir. PCA, veri setinde yer alan değişkenler arasındaki ilişkileri analiz etmek ve bu ilişkileri en iyi şekilde açıklayan yeni değişkenler oluşturmak için kullanılır.

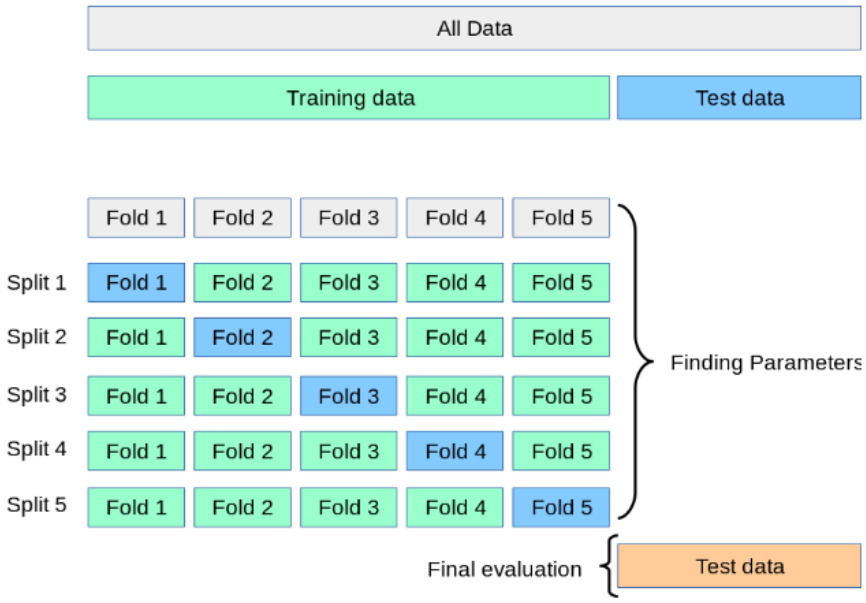
Tablo 3.1. Lasso ve PCA sonrası veri seti boyutları

	Lasso		PCA	
	Örnek Sayısı	Özellik Sayısı	Örnek Sayısı	Özellik Sayısı
Veri seti 1	5557	11	5557	7
Veri seti 2	11688	9	11688	7

3.2.3. Eğitim ve test verilerinin ayrılması

Makine öğrenmesi algoritmalarında kullanılmak üzere eğitim ve test verileri ayrıştırıldı. Bu işlem için K-Fold cross validation kullanıldı. Cross validation train-test split yaklaşımından daha sağlıklı bir yöntemdir. Çünkü tüm veri üzerinde hem test hem de eğitim işlemi yapılmasına olanak tanır. Örneğin train-test split ile %20 test, %80 eğitim olarak veriler ayrıldığında burada verilerin dağılımının nasıl gerçekleştiği bilinmediğinden bazı sapma ve hatalar oluşabilir.

K-Fold cross validation yönteminde veriseti k farklı alt kümeye bölünür. Her bir parçanın hem eğitim hem test verisi olarak kullanılması sağlanır. Bu sayede dağılımından kaynaklı oluşabilecek hatalar minimuma indirilir. Bu çalışmada k değeri 10 alınarak analizler gerçekleştirildi.



Şekil 3.3. K-Fold çalışma prensibi

3.2.4. Makine öğrenmesi algoritmaları

Çalışmada 6 farklı makine öğrenmesi algoritması kullanıldı. Hem düşük sayıda veri içeren veri seti hem de daha yüksek sayıda veri içeren veri seti bu algoritmaların her biri ile çalıştırıldı. Algoritmalarda en iyi parametrelerin belirlenmesi için GridSearchCV kullanıldı.

3.2.4.1. GridSearchCV (Grid Search Cross-Validation)

GridSearchCV makine öğrenmesi algoritmalarında hiperparametrelerin ayarlamasını gerçekleştirmek için kullanılan bir optimizasyon tekniğidir. Teknik, bir hiperparametre kombinasyonu kümesi üzerinde çapraz doğrulama kullanaran en iyi hiperparametre değerlerini bulmayı hedefler.

Hiperparametre uzayı belirlendikten sonra olası hiperparametre kombinasyonları oluşturulur. Bu hiperparametrelerin kombinasyonları belirlenerek modeller oluşturulur ve çapraz doğrulama yöntemi ile model performansı ölçülür. Hiperparametreleri deneme yanılma yöntemiyle el ile ayarlamak yerine GridSearchCV bu işlemi otomatik olarak yapar ve en iyi sonuçları verir.

3.2.4.2. Rastgele orman regresyonu

Rastgele Orman regresyon (Random Forest Regression - RFR), en iyi performans gösteren ve yaygın olarak kullanılan makine öğrenmesi algoritmalarından biridir. Rastgele orman regresyon, birden fazla karar ağacı üretir ve tüm bu karar ağaçlarının

çıktılarını tek bir sonuca ulaşmak için birleştirir. İlk olarak 1984 yılında tanıtılmıştır (Breiman, 2001). Karar ağaçlarını karar düğümleri ve yaprak düğümler oluşturur. Test fonksiyonu ile karar düğümleri örnekleri değerlendirir ve sonuçlara göre ilgili dallara iletir.

RFR, oluşturulan tüm karar ağaçlarını önyükleme tekniği yoluyla yeniden birleştirme ve karar ağacının her bir düğümünü oluşturmak için öngörücülerin rastgele seçimi ile birden çok örneğin oluşturulmasıyla karakterize edilen torbalama avantajlarını birleştiren, torbalama topluluğu tabanlı bir modeldir. Böylece tahmin performansını iyileştirir (Breiman, 1996).

3.2.4.3. K-En yakın komşu regresyonu

K-En Yakın Komşu (K-Nearest Neighbors - KNN) algoritması, sınıflandırma ve regresyon problemlerinde kullanılan bir makine öğrenme algoritmasıdır. Denetimli öğrenme yöntemlerindendir. Yani etiketlenmiş verilerle eğitilir ve yeni örnekleri sınıflandırmak veya tahminlemek için kullanılır.

KNN algoritması, örneğe en yakın k adet komşuyu bulur ve bu komşuların ortalamasını alarak tahminleme yapar. En yakın komşuları belirlemek için ise genellikle Öklid olmak üzere başka benzerlik ölçütleri de kullanılabilir (Altman, 1992).

KNN algoritması basit, anlaşılır ve uygulaması kolay olan bir algoritmadır. Modelin eğitim aşamasında zaman harcamaz; veri üzerinde doğrudan tahmin yapar. Ancak büyük veri kümelerinde yavaş çalışabilir, çünkü her tahminde tüm veri kümesiyle karşılaştırma yapması gerekebilir. Sonuçlar için hafızaya ihtiyaç duyar, çünkü tahminler gerçek zamanlı olarak yapılamaz.

3.2.4.4. Gradyan artırma regresyonu

Gradyan Artırma regresyon (Gradient Boosting Regression - GBR), birden fazla zayıf karar ağacının tahminlerini birleştirerek güçlü bir tahmin modeli oluşturan bir makine öğrenimi tekniğidir. Hem regresyon hem de sınıflandırma problemleri için kullanılan güçlü bir algoritmadır.

Gradient boosting yönteminde, zayıf karar ağaçları sırayla eğitilir ve her bir sonraki karar ağacı önceki karar ağaçlarının hatalarını düzeltmek için eğitilir (Friedman, 2001). Algoritma, genellikle regresyon problemleri için hedef değişkenin ortalama

değeri olarak seçilen bir başlangıç tahminiyle başlar ve ardışık olarak önceki modellerin hatalarını en aza indiren bir dizi model oluşturur.

3.2.4.5. AdaBoost regresyon

Adaboost regresyon, makine öğrenimi alanında kullanılan bir topluluk öğrenmesi algoritmasıdır (Freund ve Schapire, 1996). AB (Adaptive Boosting), zayıf öğrenicileri bir araya getirerek güçlü bir öğrenici oluşturmayı amaçlar. Genellikle bir dizi karar ağacı kullanarak uygulanır. Önceki adımlarda hatalı tahminler yapan göreceli zayıf modelleri güçlendirerek ilerler. Doğru tahmin yapılan karar ağacı noktalarının ağırlıklarını düşürürken, yanlış tahmin yapılanların ağırlıklarını artırır. Bu şekilde her adımda ağırlıkları ayarlanan zayıf modellerin bir kombinasyonunu oluşturur ve en iyi tahmin yeteneğine sahip bir güçlü model elde etmeyi amaçlar (Schapire, 2013).

3.2.4.6. Destek vektör regresyonu

Destek Vektör regresyonu (Support Vector Regression - SVR), destek vektör makinelerinin (SVM) regresyon problemlerine uyarlanmış bir versiyonudur. Destek Vektör Regresyonu, veri noktalarını bir hiper düzlemle böler ve bu hiper düzlem üzerinde yer alan destek vektörlerini kullanarak tahmin yapar (Vapnik, 2000). Amacı, veri noktalarının hiper düzleme mümkün olduğunca sınıflandırılmasını sağlayarak regresyon hatasını minimize etmektir (Awad ve Khanna, 2015).

Destek Vektör regresyonu, düşük boyutlu veya yüksek boyutlu veri setlerinde etkili bir şekilde çalışabilir. Özellikle, aykırı değerlere karşı dirençli olması ve çeşitli veri dağılımlarında iyi bir genelleme yeteneği göstermesi nedeniyle tercih edilir.

3.2.4.7. XGBoost regresyon

XGBoost (eXtreme Gradient Boosting), gradient boosting tabanlı bir öğrenme algoritmasıdır. Gradient Boosting çerçevesi altında, veri bilimi sorunlarını hızlı ve doğru bir şekilde çözebilen genişletilebilir paralel sınıflandırma ve regresyon ağaçları (CART) geliştirir. Bu yaklaşım, zayıf öğrenicileri (genellikle karar ağaçlarını) bir araya getirerek güçlü bir tahmin modeli oluşturur (Chen, 2016).

XGBoost, yüksek performansı, güçlü tahmin yeteneği, ölçeklenebilirlik, aykırı değerlere direnç, öznitelik önemi skorları, otomatik kesişim işlemi ve esneklik gibi özellikleriyle öne çıkar.

4. ANALİZ VE BULGULAR

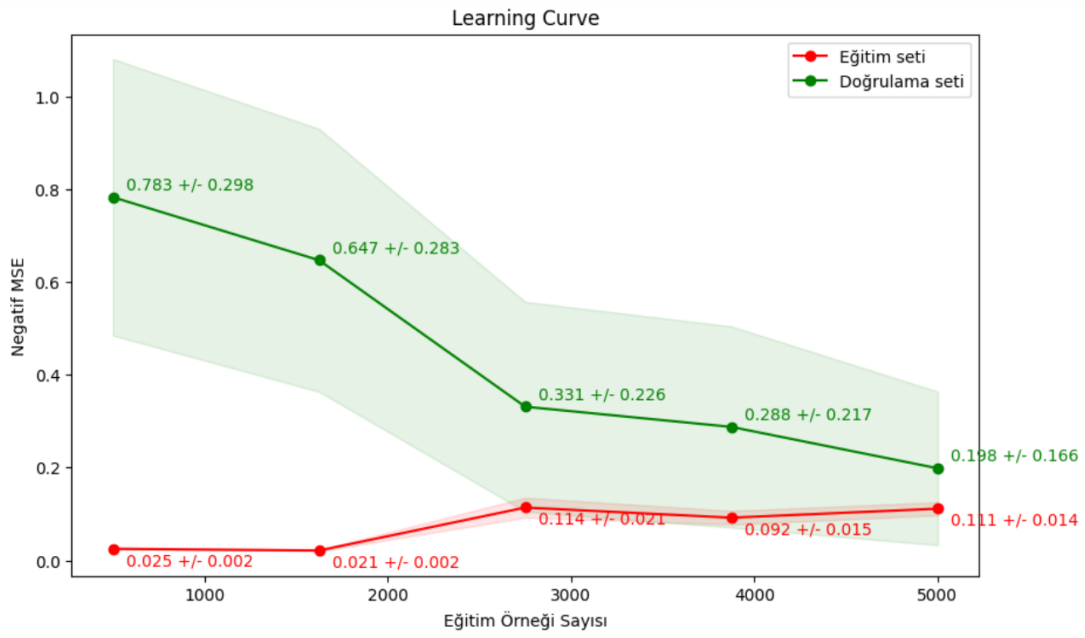
Bölüm 3'te yer alan materyal ve yöntemler kullanılarak web kazıma ile elde edilen veri setleri üzerinde makine öğrenmesi algoritmaları ile analizler gerçekleştirildi. Analiz sonuçlarında performansın gösterilmesi için learning curve grafiği kullanıldı. Learning curve, bir modelin performansının eğitim örneklerinin sayısıyla nasıl değiştiğini gösteren bir grafiktir. Eğitim örneklerinin sayısı arttıkça, modelin eğitim ve doğrulama setlerindeki performansı incelenir.

4.1. Rassal Orman Regresyonu

GridSearchCV ile random forest regresyon için en iyi parametreler belirlendi ve her iki veri seti için analiz gerçekleştirildi. Sonuçlar aşağıdaki gibidir.

- 5557 satır veri içeren veri seti ile analiz;

En iyi parametreler: {'max_depth': None, 'min_samples_split': 2, 'n_estimators': 100}



Şekil 4.1. RFR Learning Curve grafiği (Veri seti 1)

Eğitim setindeki MSE değeri (0.111) doğrulama setindeki MSE değerinden (0.198) daha düşüktür, bu da modelin eğitim verilerine daha iyi uyarlandığını gösterir. Ancak,

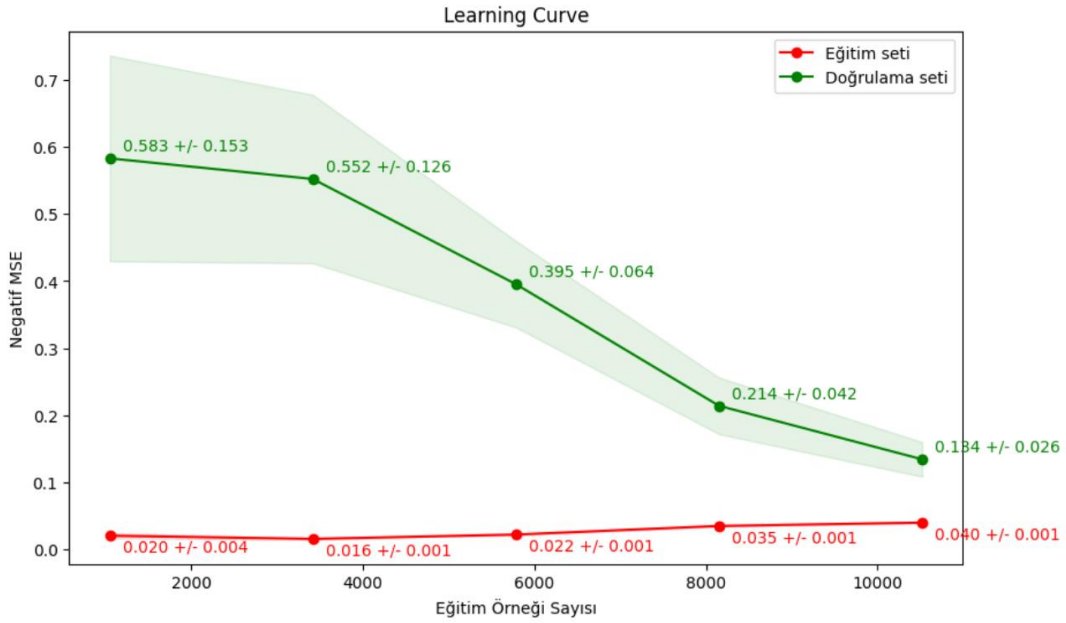
doğrulama setindeki hata daha yüksek olduğu için model, eğitim verilerine aşırı uyum yapmış olabilir.

Standart sapma değerlerine bakıldığında, eğitim setindeki hata değerleri daha homojen ve daha az değişkenken, test setindeki hata değerleri daha heterojen ve daha fazla değişkenlik göstermektedir. Bu da modelin doğrulama verilerindeki performansının eğitim verilerine göre daha değişken olabileceğine işaret eder.

Sonuç olarak, eğitim seti ve doğrulama seti arasındaki MSE değerleri ve standart sapmaları incelendiğinde, modelin eğitim verilerine iyi uyarlandığı ancak test verilerinde daha fazla hata yaptığı görülmektedir. Modelin aşırı uyum yapma eğilimi olabilir ve daha fazla genelleştirme yeteneğine ihtiyaç duyabilir.

- 11688 satır veri içeren veri seti ile analiz;

En iyi parametreler: {'max_depth': None, 'min_samples_split': 5, 'n_estimators': 100}



Şekil 4.2. RFR Learning Curve grafiği (Veri seti 2)

Bu sonuçlar, modelin aşırı uydurmadan kaçındığını gösterir. Eğitim setindeki düşük MSE değerleri, modelin eğitim verilerini iyi öğrendiğini gösterir. Aynı zamanda doğrulama setindeki MSE değerlerinin eğitim setindeki MSE değerlerinden daha yüksek olduğunu gözlenmektedir. Bu, modelin eğitim verileri üzerindeki performansının gerçek hayattaki veriler üzerindeki performansına kıyasla daha iyi olduğunu gösterir. Ancak doğrulama setindeki MSE değerleri hala düşük olduğu için, modelin yeni veriler üzerinde de iyi bir performans sergilemesi beklenir.

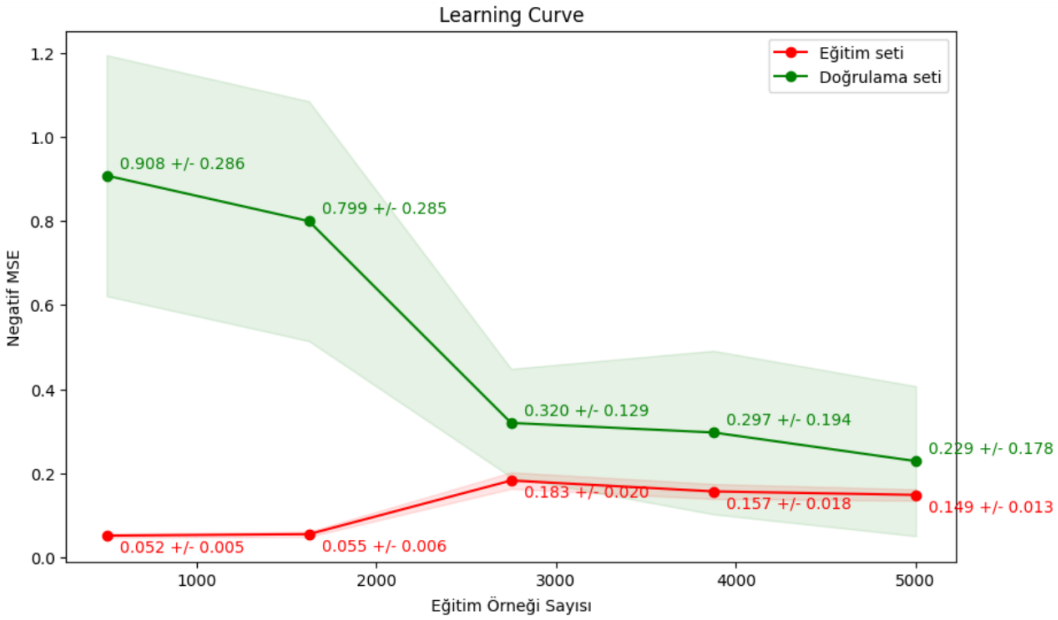
Yukarıdaki grafik ile karşılaştırıldığında modelin veri sayısı arttıkça daha iyi sonuçlar verdiği söylenebilir.

4.2. K-En Yakın Komşu Regresyon

GridSearchCV ile KNN regresyon için en iyi parametreler belirlendi ve her iki veri seti için analiz gerçekleştirildi. Sonuçlar aşağıdaki gibidir.

- 5557 satır veri içeren veri seti ile analiz;

En iyi parametreler: {'n_neighbors': 5, 'weights': 'uniform'}



Şekil 4.3. KNN Learning Curve grafiği (Veri seti 1)

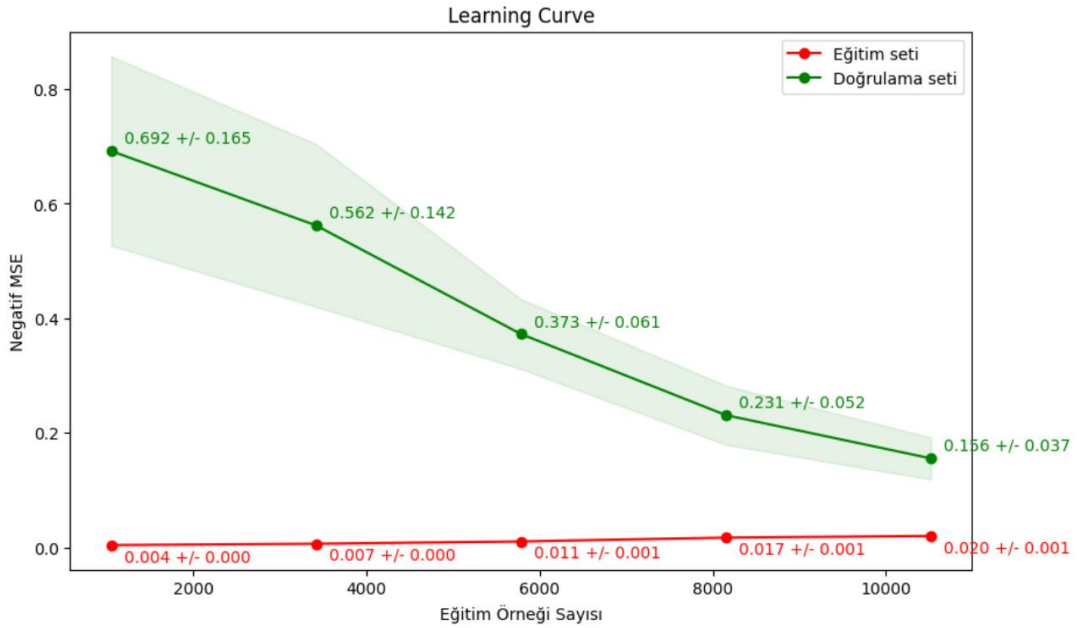
Veri boyutu arttıkça doğrulama seti MSE değeri 0.229'a düşerken, standart sapma 0.178'e düşmektedir. Bu, modelin daha fazla veriyle daha iyi bir performans sergilediğini gösterir. MSE'nin düşmesi, modelin daha iyi bir şekilde genelleştirme yapabildiğini ve tahminlerinin gerçek değerlere daha yakın olduğunu gösterir.

Eğitim setinin başlangıçtaki durumu modelin eğitim verilerine iyi uyum sağladığını ve düşük hata elde ettiğini gösterir. Ancak, veri boyutu arttıkça MSE değeri 0.149'a yükselirken, standart sapma da 0.013'e yükseliyor. Bu, modelin daha fazla veriyle birlikte eğitim setine hala iyi uyum sağladığını, ancak aşırı uyum riskinin arttığını gösterir. MSE'nin artması, modelin eğitim seti dışındaki verilere karşı daha fazla hata yapabileceğini ve daha az genelleştirici olabileceğini gösterir.

Doğrulama seti üzerindeki MSE'nin zamanla düştüğünü ve standart sapmanın azaldığını görmek olumlu bir işaret. Ancak, eğitim seti üzerindeki MSE'nin artması ve standart sapmanın yükselmesi, modelin aşırı uyum riskiyle karşı karşıya olduğunu gösterebilir.

- 11688 satır veri içeren veri seti ile analiz;

En iyi parametreler: {'n_neighbors': 7, 'weights': 'distance'}



Şekil 4.4. KNN Learning Curve grafiği (Veri seti 2)

Bu sonuçlar, modelin aşırı uyum yapmadığını ve test seti üzerinde başarısının, eğitim seti üzerindeki başarısına yakın olduğunu gösteriyor. Eğitim setinde elde edilen MSE değeri oldukça düşüktür, bu da modelin eğitim verileri üzerinde iyi bir performans gösterdiğini gösterir. Standart sapma değeri de oldukça düşüktür, bu da modelin eğitim verilerine duyarlılığının düşük olduğunu ve tahminlerin tutarlı olduğunu gösterir.

Test setinde elde edilen MSE değeri eğitim setindeki MSE değerine göre biraz daha yüksektir. Standart sapma değeri ise daha yüksektir, bu da modelin test verilerindeki performansının daha değişken olduğunu ve tahminlerin daha geniş bir hata aralığına sahip olduğunu gösterir.

Sonuç olarak veri sayısının artmasıyla birlikte modelin genel performansının iyileştiğini görülmektedir. Doğrulama seti üzerindeki MSE'nin düşmesi ve standart sapmanın azalması, modelin daha fazla veriyle birlikte daha iyi bir şekilde genelleştirme yaptığını gösterir. Eğitim seti üzerindeki MSE'nin artması ve standart

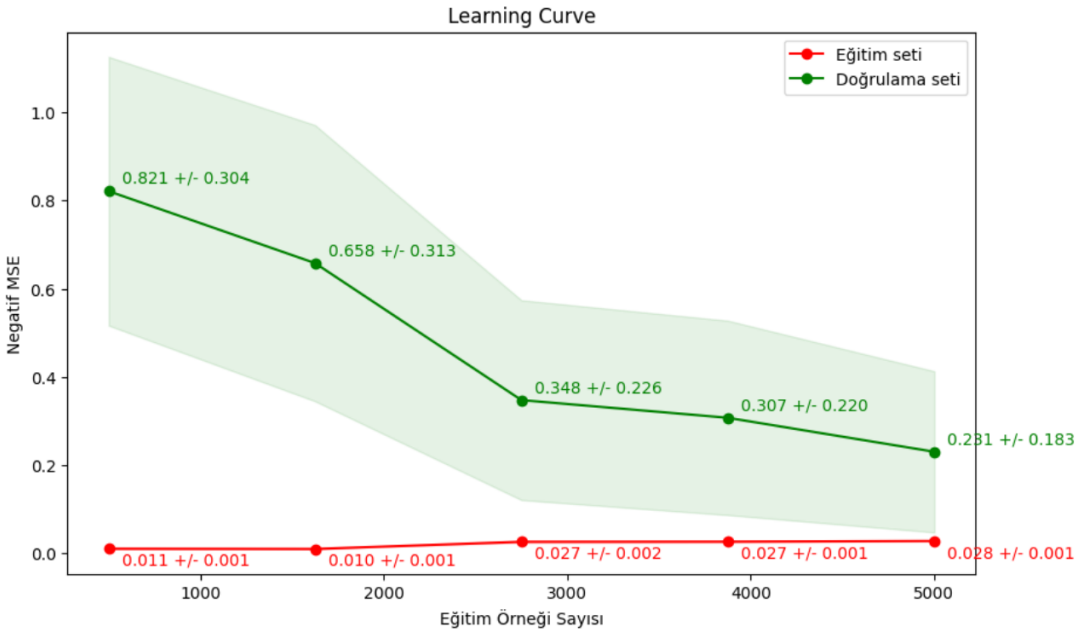
sapmanın yükselmesi, modelin aşırı uyum riskinin artabileceğine işaret edebilir ancak sonuçlar yine de olumlu görünmektedir.

4.3. Gradyan Artırma Regresyonu

GridSearchCV ile GBR için en iyi parametreler belirlendi ve her iki veri seti için analiz gerçekleştirildi. Sonuçlar aşağıdaki gibidir.

- 5557 satır veri içeren veri seti ile analiz;

En iyi parametreler: {'learning_rate': 0.1, 'max_depth': 7, 'n_estimators': 100}



Şekil 4.5. GBR Learning Curve grafiği (Veri seti 1)

Eğitim örneği sayısı arttıkça doğrulama setindeki MSE ve standart sapma değerleri azalıyor. Bu, modelin daha fazla veriyle daha iyi genelleme yaptığını gösterir. Daha fazla veriyle birlikte model doğrulama performansı artmaktadır.

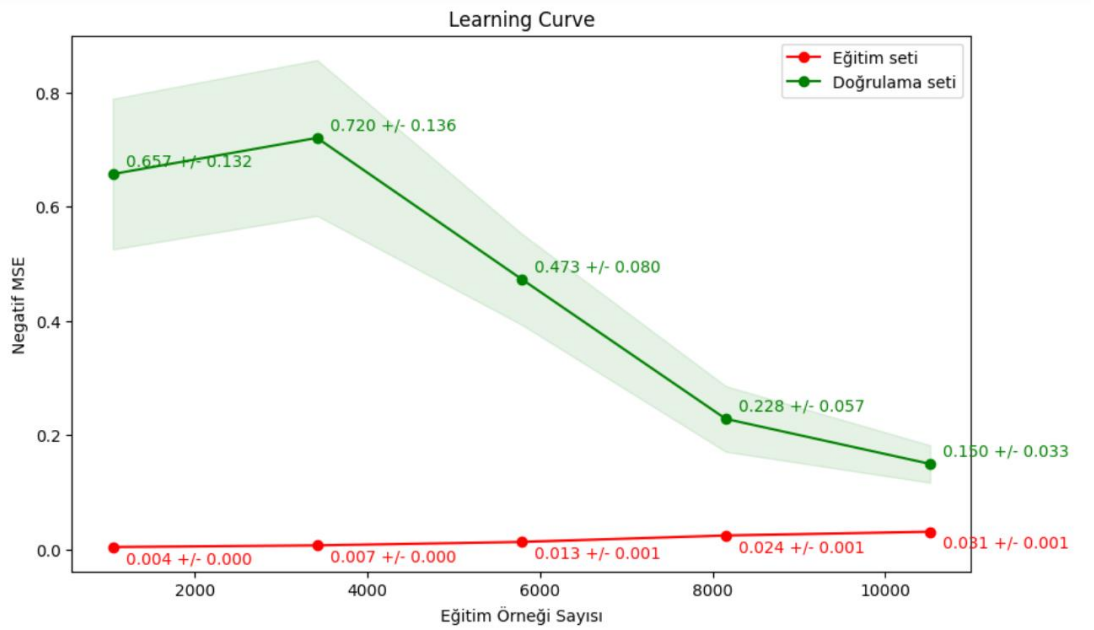
Eğitim seti için MSE değeri başlangıçta 0.011 ve standart sapma 0.001 olarak başlıyor. Veri boyutu arttıkça MSE değeri 0.028'e yükseliyor, ancak standart sapma aynı kalıyor. Bu, modelin eğitim setine daha iyi uyum sağladığını ve veri boyutunun artmasıyla birlikte modelin eğitim performansının azaldığını gösteriyor. Bu durum, modelin eğitim verilerine aşırı uyum sağladığını ve genelleme performansının düşebileceğini gösterebilir.

Genel olarak grafikte doğrulama seti performansının veri boyutuyla birlikte iyileştiği, ancak eğitim seti performansının azaldığı görülmektedir. Bu, modelin daha fazla

veriyle daha iyi genelleme yapabildiğini, ancak eğitim verilerine aşırı uyum sağlama riskini göstermektedir.

- 11688 satır veri içeren veri seti ile analiz;

En iyi parametreler: {'learning_rate': 0.1, 'max_depth': 7, 'n_estimators': 200}



Şekil 4.6. GBR Learning Curve grafiği (Veri seti 2)

Başlangıçta 0.657 MSE ve 0.132 standart sapma değerleri görülmektedir. Daha sonra, veri boyutu arttıkça MSE değeri 0.720'ye yükseliyor ve standart sapma da 0.136'ya çıkıyor. Ancak, daha fazla veriyle birlikte MSE değeri tekrar düşüyor ve veri boyutu arttıkça 0.150 MSE ve 0.033 standart sapma ile sonlanıyor. Bu durumda, modelin daha fazla veriyle daha iyi genelleme yaptığı ve MSE değerinin istikrarlı bir şekilde azaldığı gözlenmektedir.

Eğitim seti için ise MSE değeri başlangıçta 0.004 ve standart sapma 0.000 olarak başlıyor. Veri boyutu arttıkça MSE değeri 0.031'e yükseliyor, ancak standart sapma aynı kalıyor. Bu durumda da, modelin eğitim verilerine iyi uyum sağladığı ve veri boyutunun artmasıyla birlikte eğitim performansının hafifçe azaldığı görülüyor.

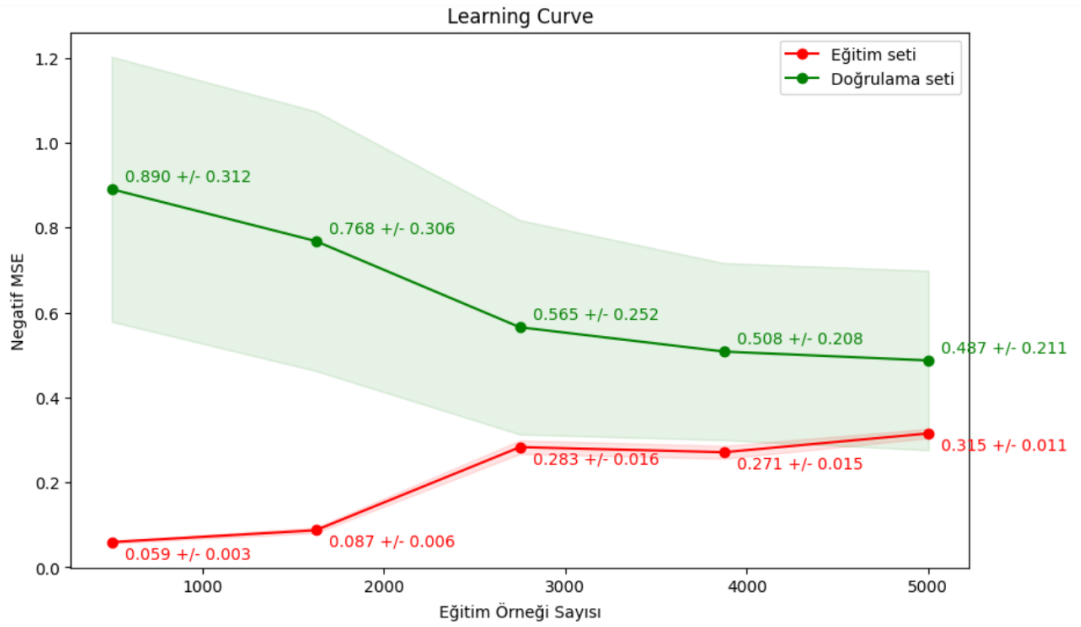
Genel olarak, grafikte doğrulama seti performansının veri boyutuyla birlikte iyileştiği, ancak eğitim seti performansının bir noktada sabitlendiği ve hatta hafifçe kötüleştiği görülmektedir. Bu, daha fazla veriyle modelin genelleme performansının artırabileceğini göstermektedir. Standart sapma değerlerine bakıldığında, daha fazla veriyle birlikte genel performansın daha da istikrarlı hale geldiği görülmektedir.

4.4. AdaBoost Regresyonu

GridSearchCV ile AdaBoost için en iyi parametreler belirlendi ve her iki veri seti için analiz gerçekleştirildi. Sonuçlar aşağıdaki gibidir.

- 5557 satır veri içeren veri seti ile analiz;

En iyi parametreler: {'learning_rate': 0.1, 'n_estimators': 50}



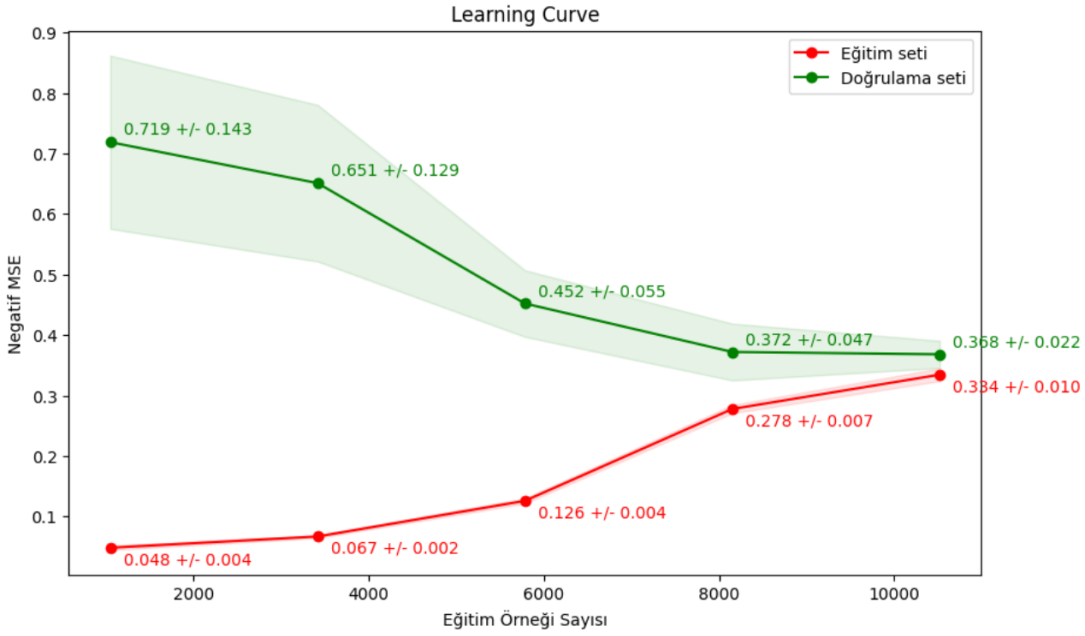
Şekil 4.7. ABR Learning Curve grafiği (Veri seti 1)

Eğitim seti ve test setindeki hataların diğer modellere göre daha yüksek çıktığı görülmektedir. Ayrıca test seti MSE'deki yüksek standart sapma, modelin tahminlerinin birbirinden oldukça farklı olduğunu ve tahminlerin güvenilirliğinin düşük olduğunu gösterir. Bu da modelin bazı veri noktalarında daha kötü performans gösterebileceğini ve tahminlerinin daha dikkatli bir şekilde değerlendirilmesi gerektiğini gösterir.

Sonuç olarak, grafiğe göre modelin eğitim setindeki MSE değeri daha iyi olsa da, test seti MSE değeri ve yüksek standart sapma, modelin genelleme yeteneğinin zayıf olduğunu, aşırı uyum (overfitting) sorunu yaşadığını göstermektedir.

- 11688 satır veri içeren veri seti ile analiz;

En iyi parametreler: {'learning_rate': 0.1, 'n_estimators': 50}



Şekil 4.8. ABR Learning Curve grafiği (Veri seti 2)

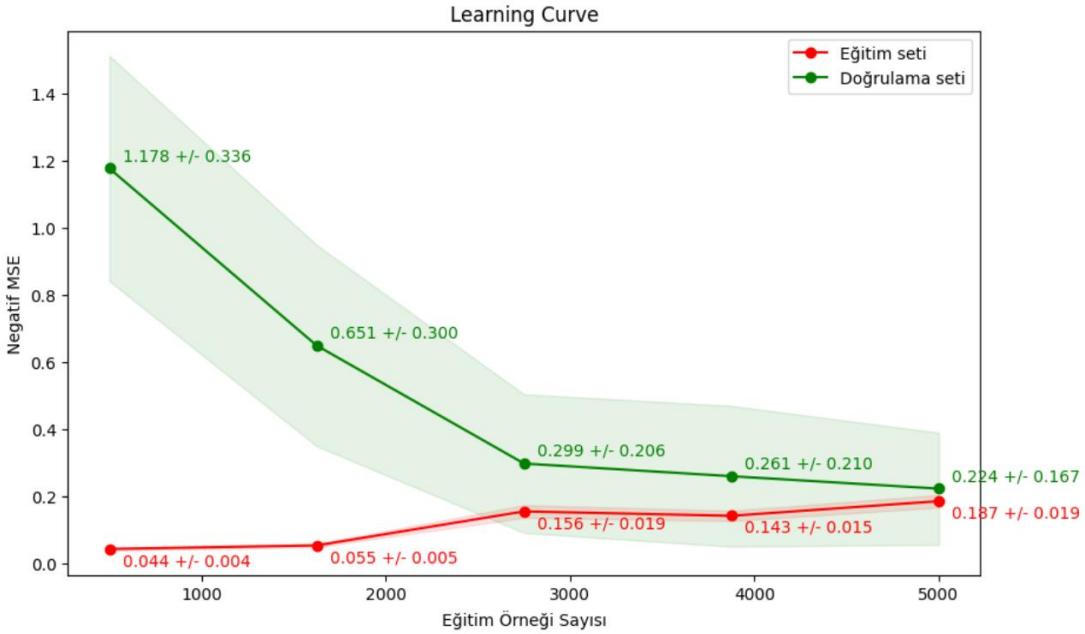
Verinin artırılması test setindeki hata oranını ve standart sapmayı düşürmüş görünüyor. Ancak grafiğin geneline bakıldığında değerler veri boyutu artarken aşırı öğrenmenin olduğuna işaret etmektedir. Daha fazla veri eklemenin bu model için faydasız olduğu söylenebilir.

4.5. Destek Vektör Regresyon

GridSearchCV ile Destek vektör regresyon için en iyi parametreler belirlendi ve her iki veri seti için analiz gerçekleştirildi. Sonuçlar aşağıdaki gibidir.

- 5557 satır veri içeren veri seti ile analiz;

En iyi parametreler: {'C': 10, 'epsilon': 0.1}



Şekil 4.9. SVR Learning Curve grafiği (Veri seti 1)

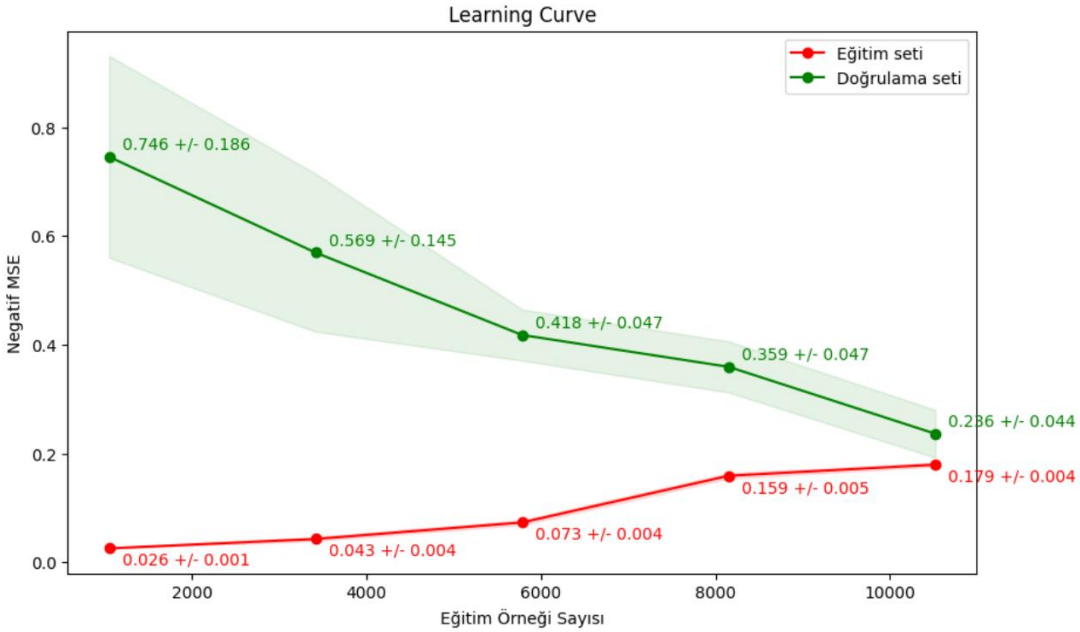
Başlangıçta, doğrulama seti MSE değeri 1.178 ve standart sapma değeri 0.336 olarak belirlenmiş. Bu durum modelin başlangıçta kötü performans gösterdiğini ve yüksek bir hata oranına sahip olduğunu gösteriyor. Ancak, veri boyutu arttıkça, MSE değeri 0.224'e düşerken standart sapma değeri 0.167'ye düşüyor. Bu, modelin daha fazla veriyle daha iyi performans gösterdiğini ve daha az varyasyona sahip olduğunu gösteriyor. Genel olarak, doğrulama seti performansının veri boyutu arttıkça iyileştiği söylenebilir.

Eğitim seti başlangıç değerleri incelendiğinde ise eğitim seti MSE değeri 0.044 ve standart sapma değeri 0.004 olarak belirlenmiş. Bu durum modelin başlangıçta iyi performans gösterdiğini ve düşük bir hata oranına sahip olduğunu gösteriyor. Ancak, veri boyutu arttıkça, MSE değeri 0.187'ye yükselirken standart sapma değeri 0.019'a yükseliyor. Bu, modelin daha fazla veriyle birlikte eğitim hatasında artış olduğunu ve daha fazla varyasyona sahip olduğunu gösteriyor. Bu durumda, eğitim setinin performansı veri boyutu arttıkça kötüleşiyor gibi görünüyor. Bu, modelin aşırı uyum yapabileceği ve genelleme yeteneğinin düşük olabileceği anlamına gelebilir.

Genel olarak, doğrulama setindeki performansın eğitim setindeki performanstan daha iyi olduğu görülmektedir. Yani modelin yeni verilerle daha iyi başa çıkmak için iyi genelleme yapabildiğini gösterir. Ancak, eğitim setindeki hata değerlerinin artması, modelin aşırı uyum yapma eğiliminde olduğunu ve daha fazla veriyle birlikte daha iyi performans göstermesinin beklenebileceğini gösterir.

- 11688 satır veri içeren veri seti ile analiz;

En iyi parametreler: {'C': 10, 'epsilon': 0.1}



Şekil 4.10. SVR Learning Curve grafiği (Veri seti 2)

Önceki çıktılara göre doğrulama seti performansının iyileştiği görülüyor. Grafik eğitim seti için ise modelin başlangıçta düşük bir hata oranına ve düşük varyasyona sahip olduğunu gösteriyor. Veri boyutu arttıkça, MSE değeri 0.179'a yükselirken standart sapma değeri 0.004'e yükseliyor. Bu, modelin daha fazla veriyle birlikte eğitim hatasında bir artış olduğunu ve daha fazla varyasyona sahip olduğunu gösteriyor. Ancak, eğitim seti performansı hala kabul edilebilir bir düzeyde görünüyor.

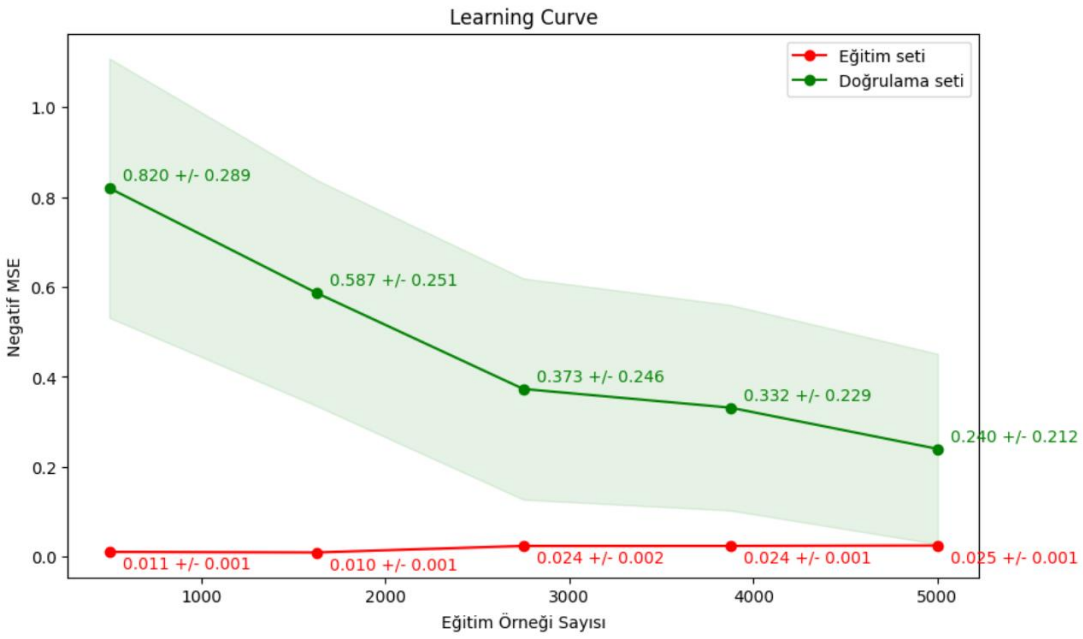
Veri boyutunu artırmanın genellikle model performansını iyileştirdiği ve daha iyi genelleme yeteneği sağladığı söylenebilir. Ancak, eğitim seti hatalarındaki artış dikkate alınmalı ve modelin aşırı uyum yapma eğiliminde olduğu için daha fazla veriye ihtiyaç duyduğu göz önünde bulundurulmalıdır.

4.6. XGBoost Regresyon

GridSearchCV ile XGBoost regresyon için en iyi parametreler belirlendi ve her iki veri seti için analiz gerçekleştirildi. Sonuçlar aşağıdaki gibidir.

- 5557 satır veri içeren veri seti ile analiz;

En iyi parametreler: {'learning_rate': 0.1, 'max_depth': 5, 'n_estimators': 500}



Şekil 4.11. XGB Learning Curve grafiği (Veri seti 1)

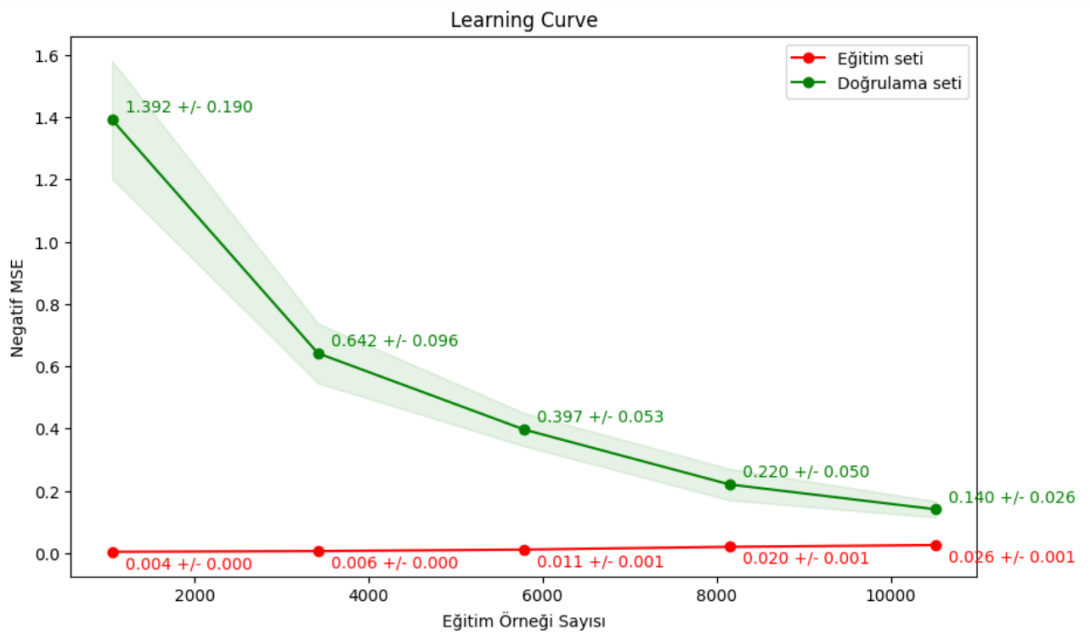
Eğitim setinde, başlangıçta düşük bir MSE değeri (0.011) ve düşük bir standart sapma (0.001) var. Bu, modelin eğitim verilerini iyi bir şekilde öğrendiğini ve düşük bir hataya sahip olduğunu gösterir. Ancak, doğrulama setinde başlangıçta daha yüksek bir MSE değeri (0.820) ve daha yüksek bir standart sapma (0.289) olduğu görülüyor. Bu, modelin eğitim verilerine aşırı uyma (overfitting) yapmış olabileceğini veya modelin genelleme yeteneğinin düşük olduğunu gösterebilir. Model, eğitim setinde iyi performans gösterirken, yeni veriler üzerinde daha yüksek hatalar yapabiliyor olabilir.

Veri boyutu arttıkça, doğrulama setindeki MSE değeri düşüyor ve standart sapma azalıyor. Bu, modelin daha fazla veriyle daha iyi performans gösterdiğini ve genelleme yeteneğinin arttığını gösterir. MSE değerinin azalması, modelin tahminlerinin doğruluğunun arttığını ve verilere daha iyi uyarlandığını gösterir.

Özet olarak, eğitim seti üzerinde model iyi performans gösterirken, doğrulama setinde başlangıçta düşük performans sergiliyor. Ancak, veri boyutu arttıkça modelin genelleme yeteneği iyileşiyor ve doğrulama setindeki hata azalıyor. Bu sonuçlar, modelin daha fazla veriyle daha iyi çalışabileceğini ve genelleme yeteneğini artırabileceğini gösteriyor.

- 11688 satır veri içeren veri seti ile analiz;

En iyi parametreler: {'learning_rate': 0.1, 'max_depth': 5, 'n_estimators': 1000}

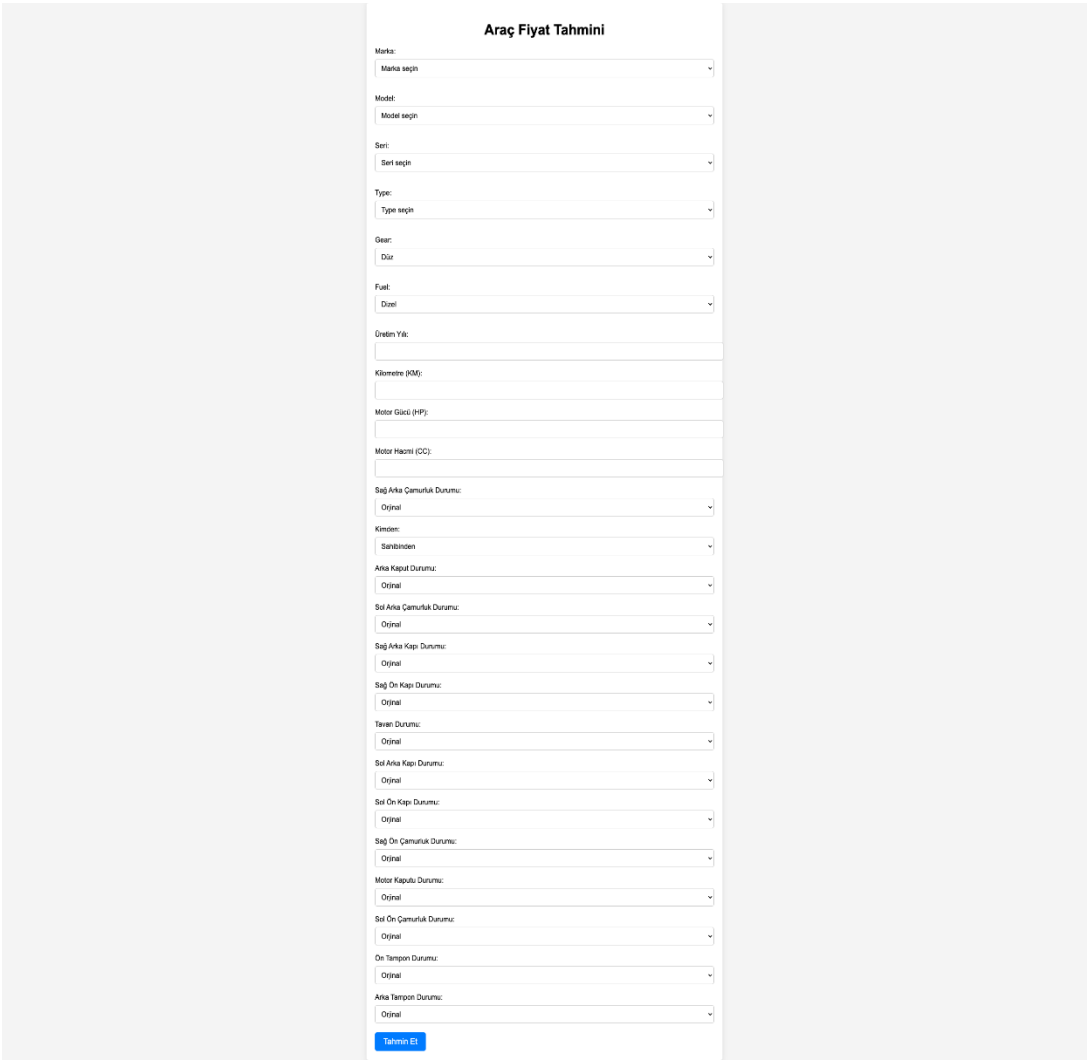


Şekil 4.12. XGB Learning Curve grafiği (Veri seti 2)

Grafik incelendiğinde model eğitim seti üzerinde iyi performans gösteriyor, ancak doğrulama seti üzerinde başlangıçta düşük performans sergiliyor. Veri sayısının artmasıyla birlikte modelin genelleme yeteneği iyileşiyor ve doğrulama setindeki hata azalıyor. Bu durumda veri boyutunu artırmanın genellikle model performansını iyileştirdiği ve daha iyi genelleme yeteneği sağladığı söylenebilir.

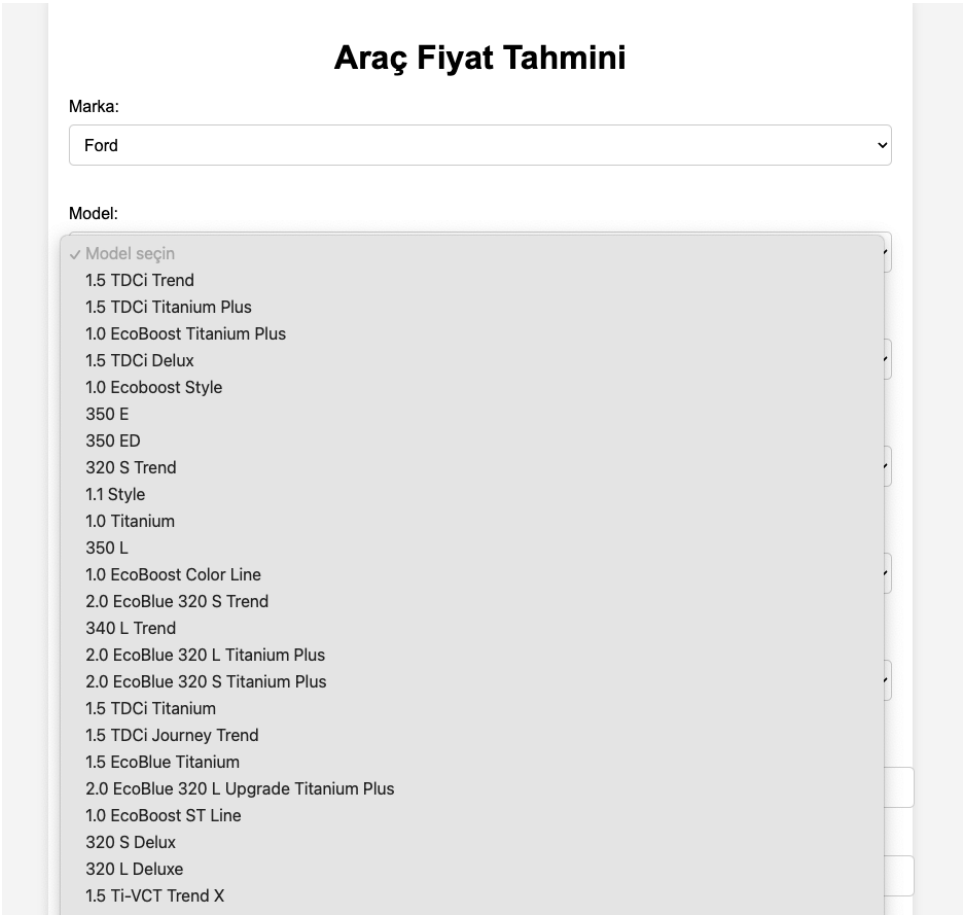
4.7. Fiyat Tahmin Arayüzü

Analiz sonrasında en iyi sonucu veren KNN algoritması kullanılarak bir arayüz tasarlandı. Kullanıcıdan girdiler alınarak veri seti 2 ile eğitilen modelle değerlendirildi ve tahmin değeri elde edildi. Şekil 4.13'te gösterildiği gibi kullanıcıdan araç özellikleri alındı. Marka, model gibi alanlar için Şekil 4.14'teki gibi açılır pencereler ile kullanıcının seçim yapabilmesi sağlandı. Seçilen markaya ait modellerin bir sonraki pencerede otomatik filtrelenmesi özelliği eklendi. Şekil 4.15, 4.16, 4.17 ve 4.18'de kullanıcıdan alınan değerler ile bir tahminleme örneği gösterildi. Ford marka, 1.5 TDCİ Delux Tourneo, camlı, düz vites, 2023 model, dizel, motor gücü 100, motor hacmi 1498, galeriden olan araç için hasar bilgileri, arka kaput boyalı, sağ ön kapı değişen, sol ön çamurluk değişen, motor kaputu değişen, kalan parçalar orjinal olarak seçildi. Tahmin değeri 487738 olarak elde edildi.

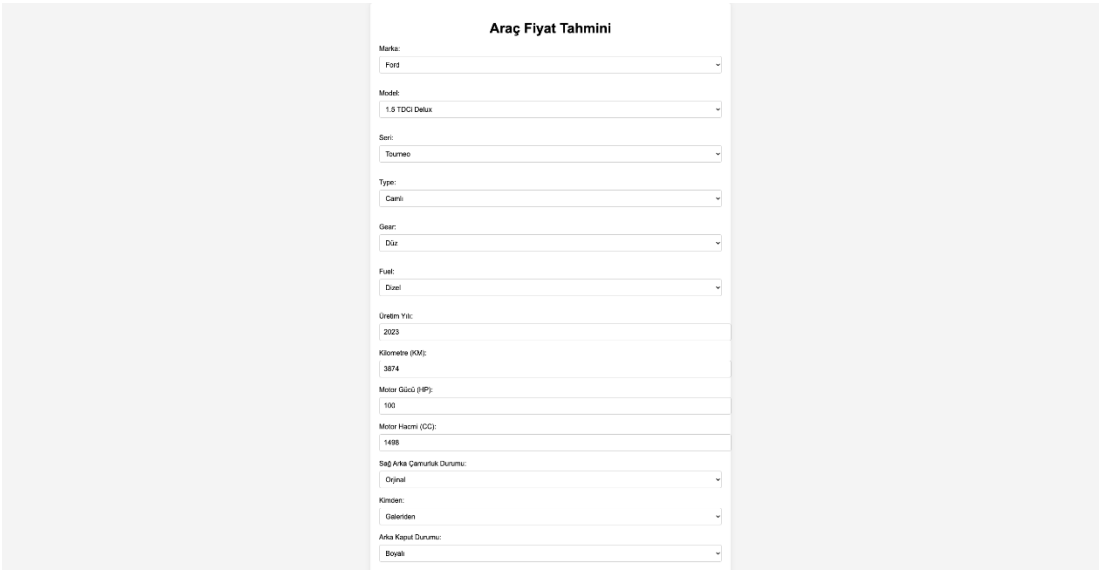


Şekil 4.13. Tahmin arayüzü

Şekil 4.14. Filtreleme örneği (marka)



Şekil 4.15. Filtreleme örneği (model)



Şekil 4.16. Fiyatı tahmin edilecek araç bilgilerinin kullanıcıdan alınması

price	brand	model	seri	type	kimden	year	fuel	gear	km	hp	cc	Sağ Arka Çamurluk	Arka Kaput	Sol Arka Çamurluk	Sağ Arka Kapı	Sağ Ön Kapı	Tavan	Sol Arka Kapı	Sol Ön Kapı	Sağ Ön Çamurluk	Motor Kaputu	Sol Ön Çamurluk	Ön Tampon	Arka Tampon
475000	Ford	1.5 TDCI Delux	Tourneo	Camlı	Galeriden	2023	Dizel	Düz	38.726	100	1498	0	2	0	0	1	0	0	0	1	1	2	0	0

Şekil 4.17. Veritabanından alınan asıl değerler

Şekil 4.18. Modelden alınan tahmin sonucu

5. TARTIŞMA VE SONUÇLAR

Makine öğrenmesi algoritmaları ile yapılan analizler sonucunda aşağıdaki çıktılar elde edilmiştir. Tabloda R^2 skorlar, MSE ve RMSE değerleri bulunmaktadır.

Tablo 5.1. Veri seti 1 ve veri seti 2 ile yapılan analizlerin sonuçları

		RFR	KNN	GBR	ABR	SVR	XGB
Veri seti 1	R^2	0.889	0.854	0.970	0.676	0.811	**0.973**
	MSE	0.110	0.145	0.029	0.323	0.188	**0.026**
	RMSE	0.333	0.382	0.172	0.568	0.434	**0.161**
Veri seti 2	R^2	0.960	**0.978**	0.967	0.658	0.820	0.973
	MSE	0.039	**0.021**	0.032	0.341	0.179	0.026
	RMSE	0.197	**0.145**	0.179	0.584	0.423	0.164

Genel çerçevede sonuçlar incelendiğinde RFR, KNN ve SVR modelleri için veri sayısı artırmanın sonuçları olumlu şekilde etkilediği, skorların arttığı ve hata değerlerinin azaldığı, diğer modeller için ise veri sayısı arttığında performansın biraz düştüğü görülmektedir.

R^2 skorlar ve hata metrikleri (MSE, RMSE) ile birlikte sonuçlar incelendiğinde AdaBoost regresyon modelinin diğer modellerden daha kötü sonuçlar verdiği çıkarılmaktadır. Bölüm 4'te yer alan grafik ile birlikte değerlendirildiğinde veri eklemenin bu modelde sonuçları iyileştirmediği görülmektedir.

Diğer modellerin sonuçları incelendiğinde 5557 satır veri içeren veri seti 1 için Gradient Boosting tabanlı GBR ve XGB modellerinin sonuçlarının kalan modellerden daha iyi olduğu görülmektedir. İki model kendi aralarında kıyaslandığında ise XGBoost modelinin daha iyi sonuçlar verdiği belirlenmiştir. 11688 satır veri içeren

veri seti 2 için sonuçlara bakıldığında ise en iyi sonuçları KNN modelinin verdiği görülmektedir. En iyi sonuçlar Tablo 5.1'de koyu renkli olarak gösterilmiştir.

Çalışma sonucunda web kazıma yöntemleri ile web sitesi üzerinden toplanan veriler makine öğrenmesi algoritmaları ile analiz edilmiş ve fiyat tahminleme için boyut olarak farklı veri setleri için en iyi sonuçları veren algoritmalar belirlenmiştir. Bazı modeller için veri sayısını artırmanın sonuçları olumlu, bazı modeller için ise olumsuz etkilediği görülmüştür.

Çalışmadan yola çıkarak farklı kaynaklardan benzer verilerin web kazıma yöntemleri ile toplanıp veri setinin genişletilebileceği ve analizler sonrası uygun model seçimi yapılarak performanslı tahminlerin yapılabileceği çıkarımı yapılabilmektedir. Geliştirilmesi halinde bu çalışma çeşitli yöntemler ile web kazıma işlemi yaparak web sitelerinden sürekli veri çekebilecek hale getirilebilir ve veri toplanabilir, toplanan veriler analiz edilerek tahmin sonuçlarını paylaşabilir.

KAYNAKLAR

Altman, N. S. (1992). An introduction to kernel and nearest-neighbor nonparametric regression. The American Statistician, vol.46, s. 175–185.

Asghar, M., Mehmood, K., Yasin, S., & Khan, Z. M. (2021). Used Cars Price Prediction using Machine Learning with Optimal Features. Pakistan Journal of Engineering and Technology, 4(2), 113-119.

Awad M. and Khanna R. (2015). Efficient Learning Machines, Apress.

Banerjee, R. (2014). Website Scraping. Happiest Minds Technologies.

Boegershausen, J., Borah, A., & Stephen, A. (2020). Fields of Gold: Web Scraping for Consumer Research. Marketing Science Institute Working Paper Series, 20-143.

Breiman, L. (1996). Bagging Predictors. Machine Learning, 24(2), 123-140.

Breiman, L. (2001). Random Forests. Machine Learning, 45(1), 5-32.

Broucke, S. and Baesens, B. (2018). Practical Web Scraping for Data Science: Best Practices and Examples with Python. Berkeley, CA, Apress.

Chen, K.P., Liang, T.P., Yin, S.Y., Chang, T., Liu, Y.C., & Yu, Y.T. (2020). How serious is shill bidding in online auctions? Evidence from ebay motors. Available at SSRN 3279292.

Chen, T., Guestrin, C. (2016). XGBoost: A scalable tree boosting system. Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 13-17-Augu, 785–794.

Cross-validation: evaluating estimator performance. https://scikit-learn.org/stable/modules/cross_validation.html adresinden 4 Mayıs 2023 tarihinde alınmıştır.

Freund, Y. and Schapire, R. E. (1996). Experiments with a new boosting algorithm. Icml'96: Proceedings of the Thirteenth International Conference on Machine Learning, 148-156.

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. Annals of Statistics, 29(5), 1189-1232.

Gegic, E.; Isakovic, B.; Keco, D.; Masetic, Z.; Kevric, J. (2019). Car price prediction using machine learning techniques. TEM J. 2019, 8, 113-118.

Haddaway, N. (2015). The use of web-scraping software in searching for grey literature. Grey Journal, 11(3), 186-190.

Henrys, K. (2021). Importance of web scraping in e-commerce and e-marketing. Available at SSRN 3769593.

Khder, M. (2021). Web Scraping or Web Crawling: State of Art, Techniques, Approaches and Application. International Journal of Advances in Soft Computing and its Applications, 13(3), 145-168.

Milev, P. (2017). Conceptual Approach for Development of Web Scraping Application for Tracking Information. Economic Alternatives, 475-485.

Oto Ekspertiz Raporunda Ne Yazar?. https://yolcu360.com/blog/oto-ekspertiz-raporunda-ne-yazar adresinden 3 Mayıs 2023 tarihinde alınmıştır.

Pandey, A., Rastogi, V., & Singh, S. (2020). Car's selling price prediction using random forest machine learning algorithm. In 5th International Conference on Next Generation Computing Technologies.

Schapire, R. E. (2013). Explaining adaboost. In Empirical Inference, pp. 37–52.

Vapnik V. (2000). The Nature of statistical Learning Theory. Springer, New York.

ÖZGEÇMİŞ

Ad-Soyad : Seda YILMAZ

ÖĞRENİM DURUMU:

- **Lisans :** 2019, Sakarya Üniversitesi, Bilgisayar ve Bilişim Bilimleri Fakültesi, Bilgisayar Mühendisliği
- **Lisans :** 2020, Sakarya Üniversitesi, Mühendislik Fakültesi, Endüstri Mühendisliği (Çift Anadal)
- **Yükseklisans :** 2023, Sakarya Üniversitesi, Fen Bilimleri Enstitüsü, Bilişim Sistemleri Mühendisliği

MESLEKİ DENEYİM VE ÖDÜLLER:

- Ağustos 2020 tarihinden beri telekomünikasyon yazılım çözümleri sunan bir firmada Yazılım Test Mühendisi olarak çalışıyor.