YBS Ansiklopedi

www.YBSAnsiklopedi.com

Cilt 12, Sayı 2, Haziran 2024

ARAÇ KİRALAMA ŞİRKETİ VERİ KÜMESİ İLE MALİYET TAHMİNİ

Emrullah KARACAN

1. Düzce Üniversitesi Bilgisayar Mühendisliği Bölümü, emrullah.karacan33@gmail.com

ÖZET

Bu makalede gelişen ve büyüyen dünyada değeri gittikçe artan verinin maliyet analizi sürecinde kullanılmasının önemine değinilmiş ve günümüzde neden ihtiyaç duyulduğuna ve şirketlere ne gibi faydalar sağladığından bahsedilmiştir. Yapılan çalışmada bir araç kiralama şirketinin bakım maliyeti verileri kullanılmıştır. Bu veriler üzerinde önişleme, veri analizi ve görselleştirme işlemleri yapılmıştır. Literatür taraması sonucu Support Vector Machine, Random Forest, XGBoost, Polynomial Regression, LightGBM ve CatBoost modelleri kullanılmasına karar verilmiştir. Çalışma sonunda bulunan başarı değerleri literatürde bulunan değerlerle karşılaştırılmıştır. Bulunan sonuçlarda CatBoost modeli 0,999380 R^2 değeri ile en iyi sonucu veren model olmuştur. Bu çalışmanın temel amacı firmaların maliyetlerini en optimum seviyeye getirilmesinde yardımcı olmaktır. Optimum düzeyde maliyetler firmalar kâr etmesini sağlarken müşteriler uygun fiyattan ürünlerin alınabilmesini kolaylaştırmaktadır. Böylece bu çalışma ekonomik büyümeye fayda sağlamaktadır.

Anahtar Kavramlar: Maliyet Analizi, Makine Öğrenmesi, Veri Bilimi, Maliyet Tahmini

SUMMARY

This article emphasizes the increasing importance of data in the cost analysis process in the evolving and growing world, discussing why it is essential today and the benefits it provides to companies. The study focuses on the maintenance cost data of a car rental company. Preprocessing, data analysis, and visualization processes have been performed on these data.Following a literature review, it was decided to employ Support Vector Machine, Random Forest, XGBoost, Polynomial Regression, LightGBM, and CatBoost models. The success metrics obtained at the end of the study were compared with those found in the literature. Among the results, the CatBoost model stood out as the most successful, achieving an R^2 value of 0.999380. The primary objective of this study is to assist companies in optimizing their costs. Achieving costs at an optimum level not only enables companies to generate profits but also facilitates customers in purchasing products at affordable prices. Therefore, this study contributes to economic growth by promoting both profitability for businesses and accessibility of products for consumers at reasonable prices.

Keywords: Cost Analysis, Machine Learning, Data Science, Cost Estimation

1.GİRİŞ

Dijital dönüşümün hızlanması ile günümüzde mevcut veri miktarı gittikçe artmaktadır. Veri kaynaklarının artış göstermesi ile hemen her sektör, rekabette avantaj sağlamak için verileri daha etkili bir biçimde kullanmaya çalışmaktadır [1]. Şirketler veri bilimini kullanarak sonraki çeyrekler için bütçesini hazırlama ve planlama yapma şansını da yakalamaktadır. Bu nedenle maliyet analizi, bir şirketin faaliyetlerinin devamlılığını sağlama, rekabet avantajı elde etme ve kar oranını artırma sürecinde önemli bir rol oynamaktadır [2]. Günümüzde birçok şirket harcamalarını düzgün planlayamadığı için zarar etmekte hatta bu sorunlar şirketlerin iflasına bile neden olmaktadır. Özellikle şirketlerin harcamaları düşünüldüğünde veri biliminin alt dalı olan makine öğrenmesi maliyet tahmini ve optimizasyon yaparak şirket harcamalarını önemli oranda düşürebilir [3].

Araç bakımı, araç kiralama şirketleri için önemli bir giderdir. "Araştırmalara göre bakım maliyetleri şirketlerin genel maliyetinin yaklaşık %15 ile %70'ini oluşturabilmektedir." [4]. Araç kiralama şirketindeki araç başına bakım maliyetlerindeki çok az miktardaki düşüşler bile şirketlerin bütçesi için ciddi tasarruflar sağlayabilir. Bu nedenle araç kiralama şirketleri özelinde düşünüldüğünde bakım masraflarını etkili bir şekilde analiz etmesi ve yönetmesi elzemdir [5]. Çünkü bu masraflar, şirketin genel maliyet yapısını belirleyen kritik unsurlardan biridir. Bu unsurlar düşünüldüğünde maliyet analizini iyi yapan ve alanında zirveye yükselen şirketler bulunmaktadır. Örneğin Amazon şüphesiz e-ticaret alanında bir dünya devlerinden biridir. Amazon büyük denetleme ve analiz ekibi sayesinde etkin maliyet analizleri yaparak lojistik harcamalardaki maliyet optimizasyonu ve düzenli envanter yönetimi ile harcamalarını düşürerek yıllardır büyük kar oranları açıklayan şirketler arasındadır [6]. Otomotiv sektöründe dev şirketlerden olan Toyota, dünyada birçok şirketin üretim sürecine rol model olmuştur [7]. Toyota Üretim Sistemi

(TPS) isimli sistemin amacı maliyeti minimum, verimliliği de maksimuma çekmektir [7]. Bu çalışmalarının katkısı ile Toyota yıllardır otomotiv sektöründe en başarılı şirketler arasındadır.

Wubeshet Woldemariam (2020) çalışmasıında ABD'nin kırsal eyaletleri arasındaki veriler kullanılarak yıllık otoyol bakım maliyeti tahmin etmek için Destek Vektör Makinesi (Support Vector Machine, SVM) modelini kullanmıştır [8]. Çalışmada geliştirilen modelin güvenilir tahminler yaptığı görülüyor ve R^2 değeri 0,923 ortalama kare hatası ise 1,28 olarak görülmektedir [8]. Bu çalışmada destek vektör makine modeli iyi sonuçlar verdiğinden ve ayrıca literatürde de çok değişkenli veri kümeleri üzerine bazı çalışmalarda iyi maliyet tahmini yaptığı da görülmektedir [8, 9]. Bundan dolayı bu çalışmada da destek vektör makinesi kullanılacaktır.

Haytham H. Elmousalami (2021) çalışması ise biraz farklılık göstersede maliyet analizi yapılmış bir çalışmadır. Çalışmada fiber kanal arası veri aktarımı yapan bir protokolün maliyet hesabı yapılmıştır [10]. 20 model denenmiş ve bu 20 yapay zekâ modelinden en doğru sonuç alınan modelin Aşırı Gradyan Artırma (Extreme Gradient Boosting, XGBoost) olduğu görülmüştür. Maliyet tahmininde XGBoost modelinin mutlak yüzde hatası %9,091, R^2 değeri 0,929 çıkmıştır [10]. Ayrıca literatür taranmış ve gelişmiş ağaç modellerinin benim kullanacağım veri kümesi gibi karmaşık veri kümelerinde de etkili olabileceği görülmüş bundan dolayı da bu makalede XGBoost dahil gelişmiş ağaç modelleri de kullanılacaktır [11, 12].

Sasanka Katreddi ve çalışma arkadaşlarının çalışmasında araçların bakım verileri üzerinden maliyet analizi yapılmaktadır [12]. Bu makalede yapılacak çalışmadan farkı sadece otobüs kamyon gibi ağır vasıta araçlar için bakım maliyet analizi yapılmasıdır. Aksine bu makalede kullanılacak veri kümesindeki veriler bir araç kiralama şirketinin otomobillerinin verileridir. Ağır vasıta araçların maliyet analizi için Rassal Ormanlar (Random Forest, RF) modeli kullanılmıştır. Modelin ortalama değerleri şöyledir: R^2 değeri 0,966 ve ortalama kare hatası da %4,01'dir [12].

Muhammad Kamran (2021) çalışmasında regresyon problemi için Kategori Artırma (Category Boosting, Catboost) modeli kullanılarak kireçtaşı ocağındaki sırt kırılımı tahmini yapılmıştır. Benzer yapıda hedef değişken tahmini yapılması ve regresyon problemlerinde sıkça kullanılması nedeniyle Catboost modeli bizim çalışmamızda da kullanılacak ve sonuçlar karşılaştırılacaktır. Sırt kırılımı tahmini için yapılan çalışmada R^2 değeri 0,972 ve ortalama kare hatası ise %8'dir [13].

Günümüzde veri bilimi artık birçok sektörde kullanımı hızla artmaktadır [14]. Maliyet analizi firmaların gelir gider dengesini daha doğru yönetmelerine ve bu süreçte ileride atılacak adımları daha bilinçli bir şekilde atmalarına yardımcı olur [5].

2. VERİ KÜMESİ VE ANALİZİ

Kullanılacak veri kümesi Filo Bakım Veri Kümesi (Fleet Maintenance Dataset) [15]. Bu veri kümesi bir araç kiralama şirketinin araç bakım kayıtlarını barındıran bir veri kümesidir. Veri analizi, her başarılı projenin temelini oluşturan önemli bir basamaktır [2]. Bu nedenle de modelleme ve modeli optimize etme adımlarından önce veriyi anlama ve hazırlama süreçleri bir proje için çok önemlidir [16]. Veriyi modellere vermeden önce hangi model kullanılıyorsa ona uygun şekilde hazırlanmalıdır.

2.1 Veriyi Açıklama

Bu projede 2. Bölümde de bahsedilen kaggle platformundaki Araç Filosu Bakımı Veri Seti (Fleet Maintenance Dataset) isimli veri kümesi kullanılmıştır [15].

Bu veri kümesinde 2018-05-17 ve 2023-02-19 tarihleri arası toplamda 242.269 satır ve 25 sütun dan oluşmaktadır. Veri kümesindeki değişkenler, açıklamaları ve tipleri tablo-1' de verilmiştir.

Tablo 1: Araç Filosu Bakımı Veri Seti (Fleet Maintenance Dataset) Öznitelikleri

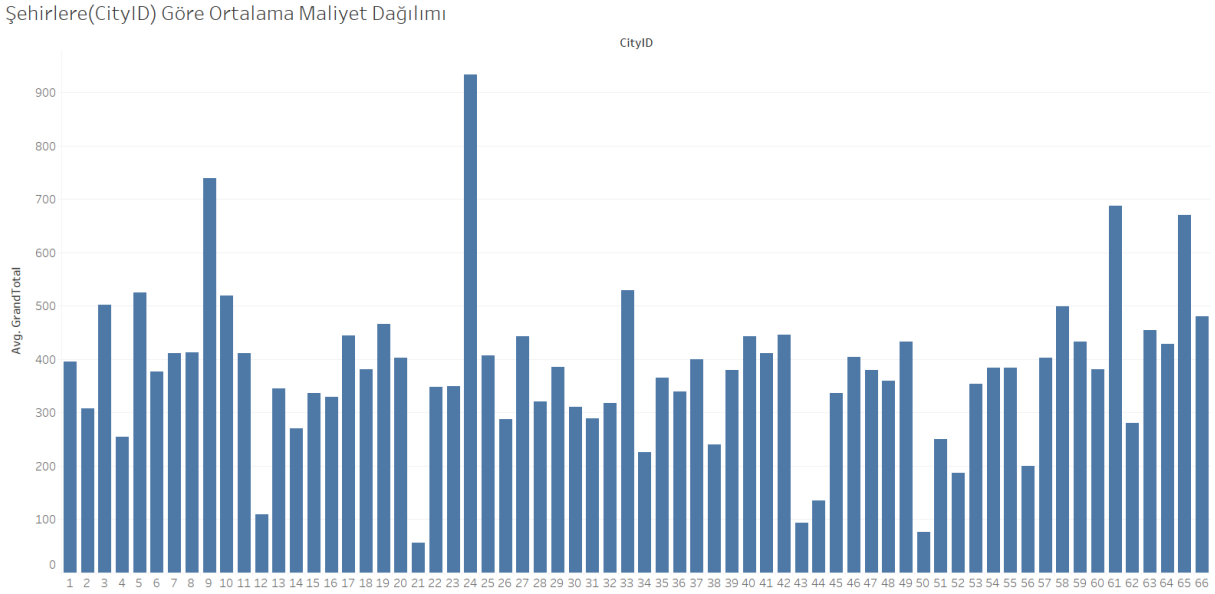
Değişkenler	Açıklama	Veri Tipi
Unnamed: 0	İndex numarası	int64
CARID	Araba kimliği	int64
Brand	Araba markası	object
Model	Araba modeli	object
SubModel	Marka ve modele göre alt model	object
Segment	Araba segmenti	object
Year	Arabanın model yılı	int64
FuelType	Arabanın yakıt tipi	object
Gear	Arabanın vites tipi	object
Milage	Bakım için beklenen kilometre değeri	int64
MaintenanceMilage	Anlık kilometre değeri	int64
CallID	Her bakım çağrısı için benzersiz kimlik	int64
StockID	Bakım için kullanılan parça ve hizmet kimliği	int64
Quantity	Kullanılan parça ve hizmet miktarı	float64
UnitPrice	Parçaların birim fiyatı	float64
SubTotal	Alt toplam (Quantity * StockID)	float64
Discount	İndirim tutarı	float64

Tax	Vergi tutarı	float64
GrandTotal	Genel Toplam (Hedef Değişken)	float64
InvoiceNo	Fatura No	int64
OperationID	Hizmet kimliği	int64
InvoiceDate	Fatura tarihi	object
VendorID	Bakım hizmeti sağlayıcısı benzersiz kimliği	int64
CityID	Bakım yapılan şehir için eşsiz kimlik	int64
MaintenanceType	Bakım türü	int64

2.2 Veri Analizi

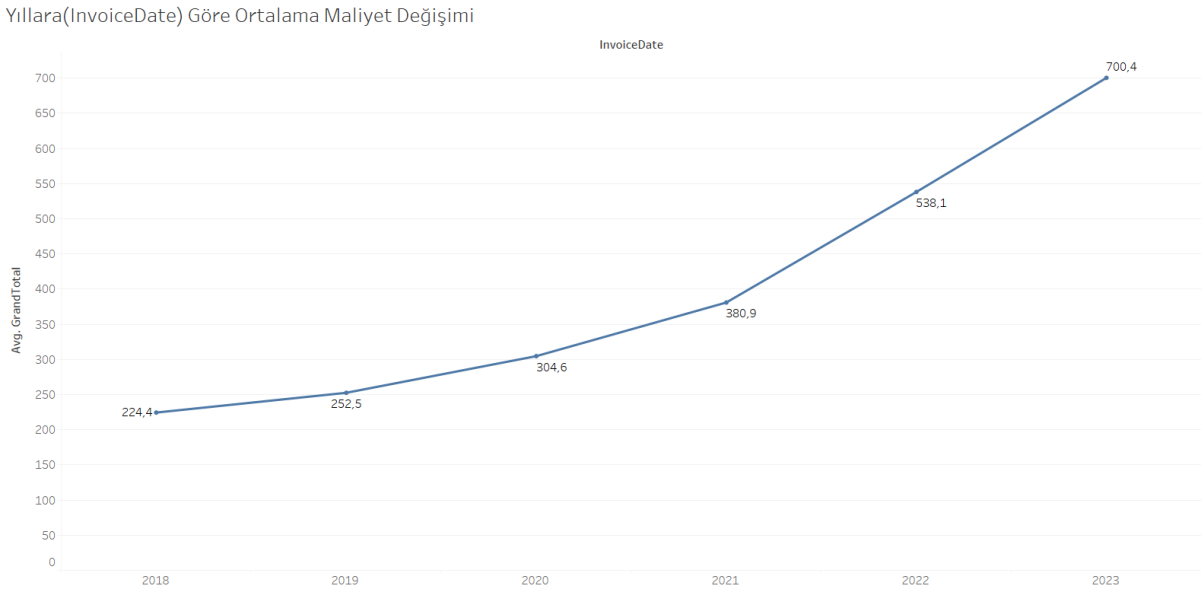
Veri analizi, verilerin anlamlı bilgilere dönüştürülmesi, incelenmesi ve sonuçlarının kullanılmasını içeren kapsamlı bir süreçtir [17]. Günümüzde veri analizi büyük bir öneme sahiptir ve çeşitli sektörlerde hayati bir rol oynamaktadır [16]. Bu süreç, daha iyi kararlar almayı sağlar [16].

CityID sütunu şehirler için plaka kodunu vermektedir yani bakım yapılan şehirler için bir eşsiz kimliktir. Şekil 1'de ise bakım yapılan şehirlerdeki ortalama maliyet dağılımı verilmiştir. Grafikten anlaşılıyor ki bakımın hangi şehirde yapıldığı maliyeti etkilemektedir. Grafikten örnek verecek olursak CityID değeri 21 ve 24 olan şehirlerdeki maliyet ortalamalarına bakıldığında bariz bir fark görülmektedir.



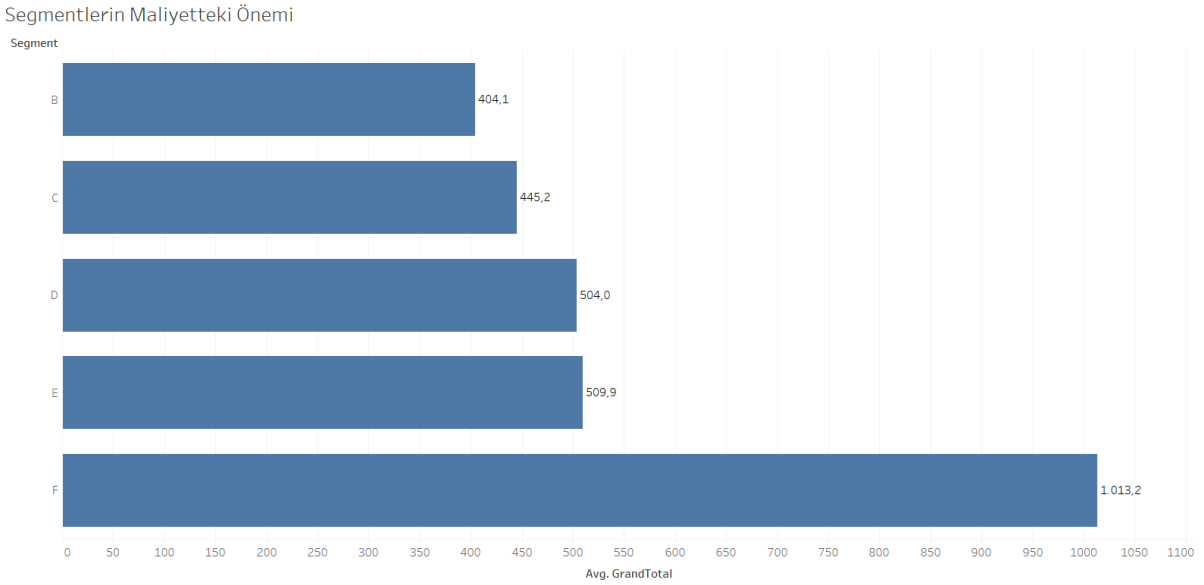
Şekil 1: Şehirlere Göre Ortalama Maliyet

Şekil 2' de Fatura Tarihi (InvoiceDate) ve Toplam Maliyet (GrandTotal) değişkenleri kullanılarak yıllara göre ortalama maliyet değişimi görülmektedir. Grafikte ortalama maliyet değerinin bariz bir şekilde yükseldiği görülmektedir. Bu nedenle etkili olabileceği düşünüldüğünden veri kümesi tarihe göre sıralanmış ve makine öğrenmesi algoritmalarında kullanılabilmesi için tarih değişkeninden yıl değeri kullanılmak üzere çekilip veri kümesine eklenmiştir.



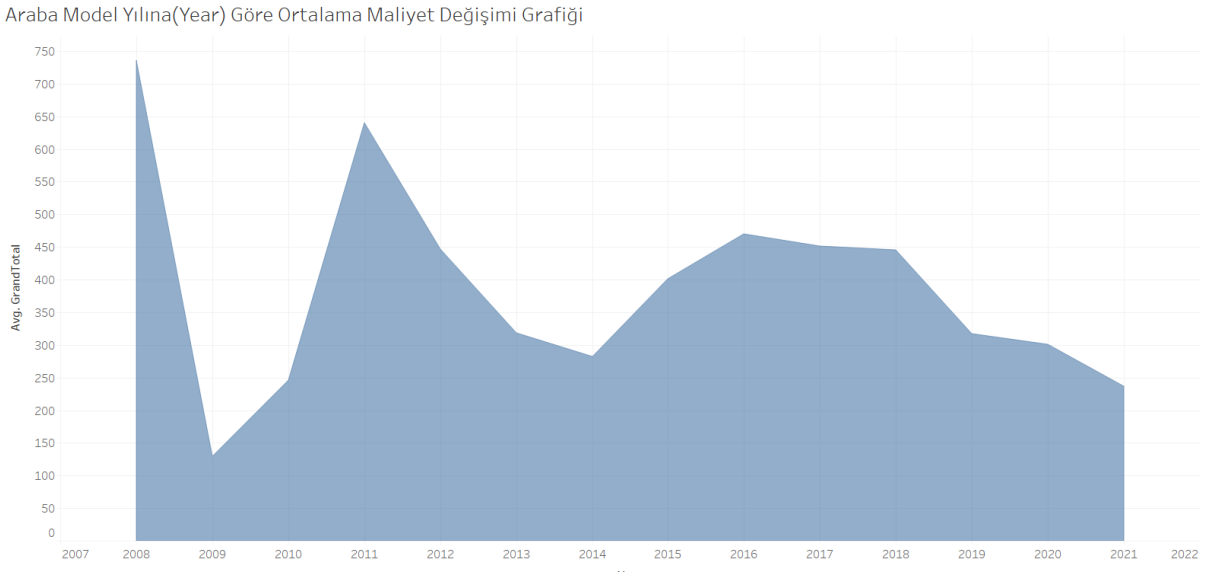
Şekil 2: Yıllara (InvoiceDate) Göre Ortalama Maliyet Değişim Grafiği

Şekil 3'te araba segmentlerinin maliyetteki etkisi bariz bir şekilde görülmektedir. Bunun için geçiş grafiği kullanılmıştır. B segmentinden F segmentine doğru koyulaşma, maliyet değerinin F segmentine doğru arttığını göstermektedir.



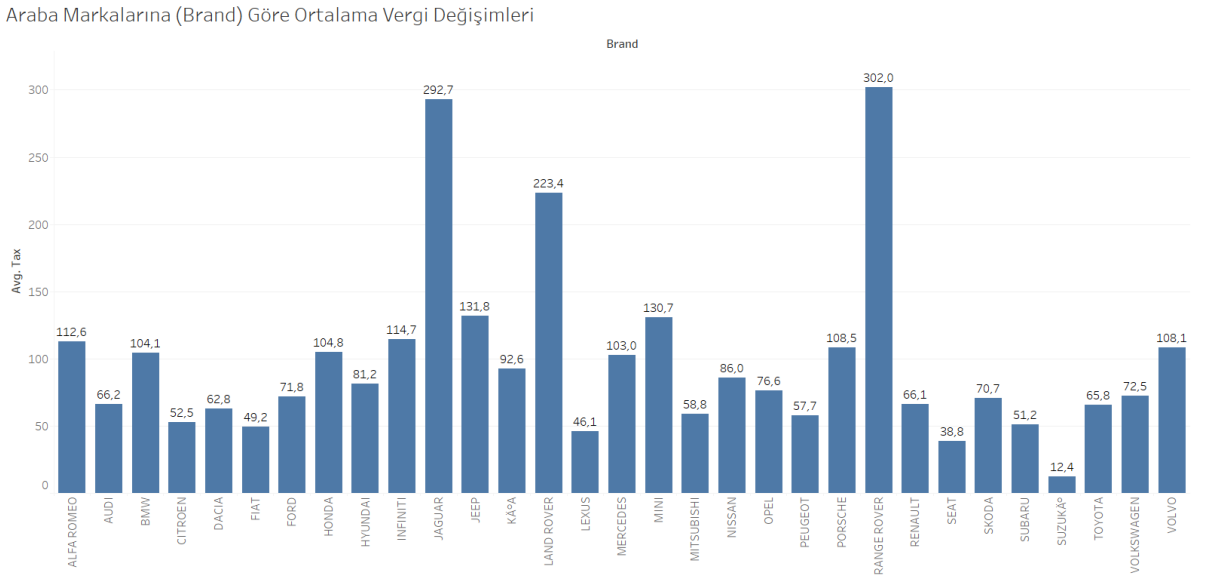
Şekil 3: Segmente göre ortalama maliyet değişimi

Şekil 4' te araba model yılının maliyete etkisi görülmektedir. Burada yaygın kanı aracın ne kadar yeni ise bakım maliyetinin de o kadar artacağı yönündedir. Ancak alan grafiğinden de görüldüğü üzere eski model araçlarda ortalama maliyet değeri en fazla olabileceği gibi çok az da olabildiği görülmektedir.



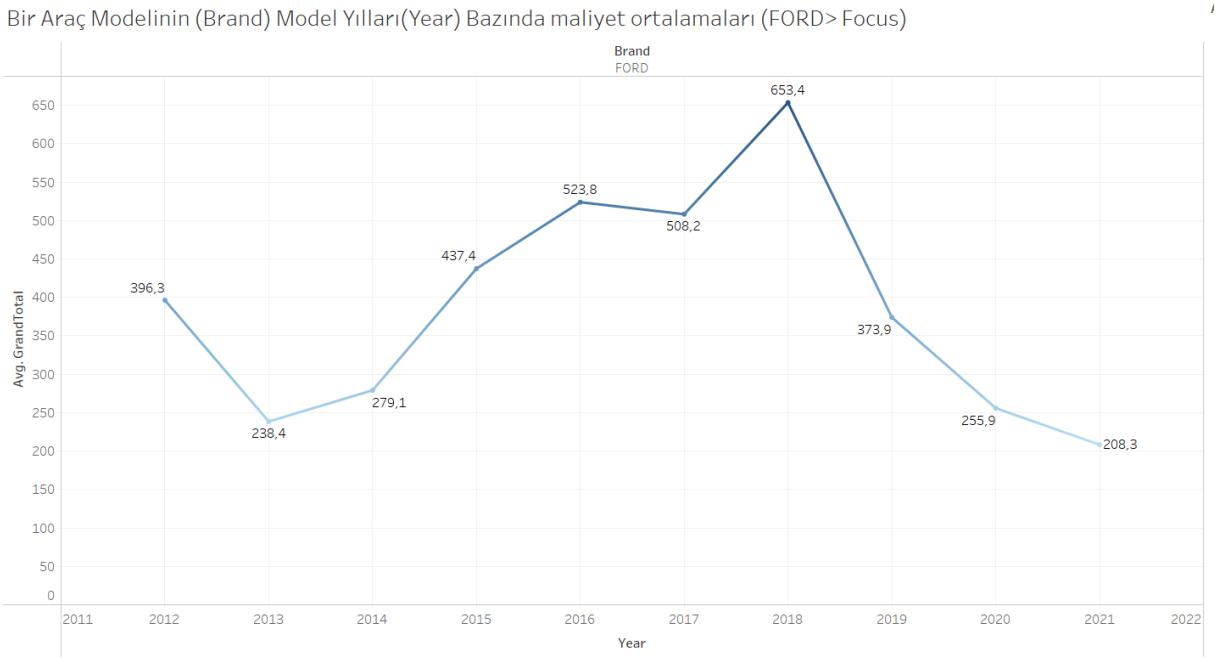
Şekil 4: Araba model yılına göre ortalama maliyet değişimi

Şekil 5'te araba markalarına göre vergi ortalaması görülmektedir. Buradan çıkarılacak sonuca göre vergi sütunu maliyet için bazı araba markalında ciddi bir etkiye sahip olmaktadır. Örnek verecek olursak Range Rover ve Jaguar marka arabalarda vergi değerleri fazla olduğu için bu arabaların bakım maliyeti de diğer arabalara göre daha yüksek değere sahip olabilir.



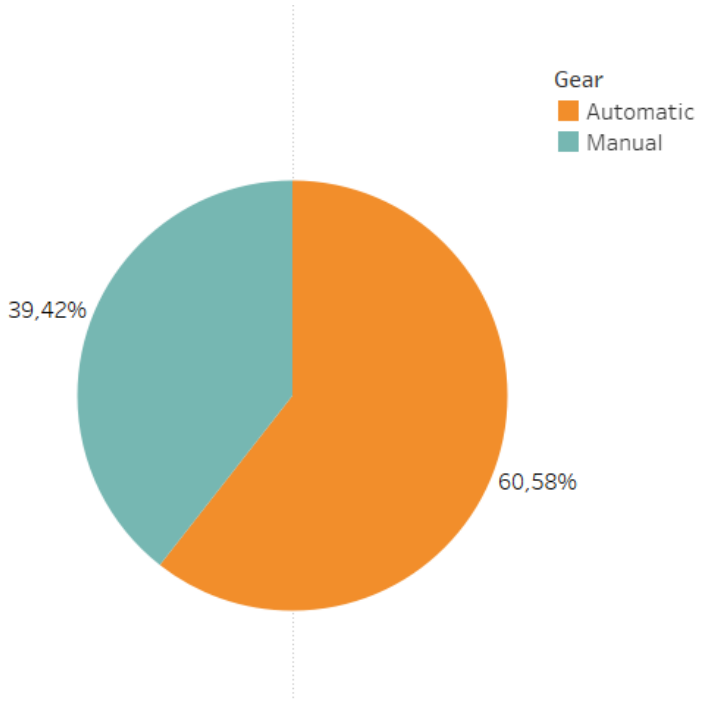
Şekil 5: Markalara göre vergi ortalaması

Şekil 6' da bir araba modelinin kendi model içinde model yılının etkili olabileceği düşünülerek çizdirilen grafik görülmektedir. Bu grafikte Ford markasının Focus modeli baz alınmış olup model yıllarının ortalama maliyet değerlerinin farklılıkları görülmektedir.



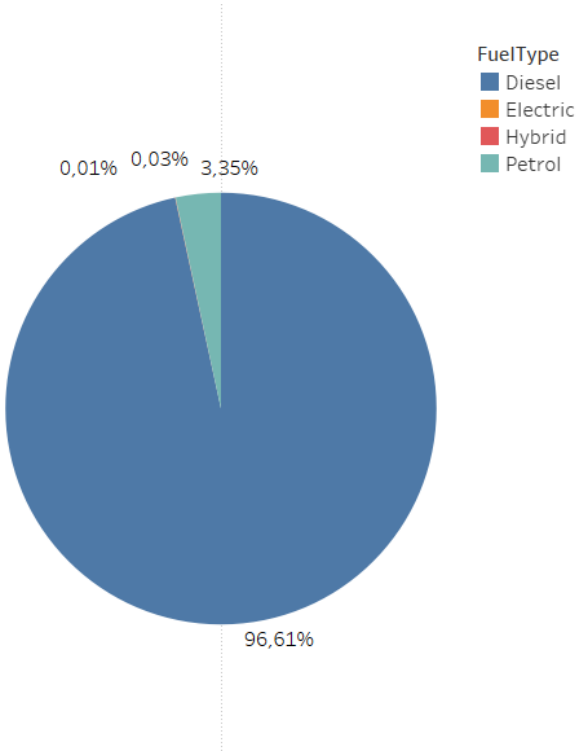
Şekil 6: Bir araba modelinin araba model yılı bazında maliyet ortalamaları

Şekil 7' de Fleet Maintenance Dataset veri kümesinde arabaların vites tipi dağılımı görülmektedir. Arabaların çoğunluğunun otomatik vites tipinde olduğu anlaşılmaktadır.



Şekil 7: Araba vites tipi dağılımı

Şekil 8' de veri kümesindeki arabaların büyük bir çoğunluğunun dizel olduğu görülmekte ve bu sütunda diğer değerlerde veri sayısının azlığından dolayı bu sütunun kullanılmamasına karar verilmiştir.



Şekil 8: Araba yakıt tipi dağılımı

2.3 Veriyi Hazırlama

Bu aşamada veri kümesindeki "InvoiceDate" adlı fatura tarihini veren sütun object tipindedir. Tarih verisi barındıran bu değişkenin tipi daha sonra kullanılabilmek için datetime tipine dönüştürülmüştür. Veri kümesindeki "Unnamed: 0", "CallID", "InvoiceNo" ve "OperationID" isimli benzersiz ID değerleri içeren sütunlar gereksiz olduğu ve model sonuçlarına olumlu katkısı olmayacağı düşünüldüğünden silinmiştir. "OperationID" isimli sütunun silinmesinin bir başka sebebi de içerisinde bütün satırlarda sadece sıfır değeri olmasıdır. Veri kümesindeki birbiri ile ilişkisi yüksek kolanlar bazen aşırı öğrenme 'ye neden olabilmektedir [13]. Aşırı öğrenme ihtimalinden dolayı "StockID" ve "Quantity" sütunlarının çarpım sonucunu barındıran "SubTotal" sütunu etkisi ölçülmüştür. İlk başta iyi sonuçlar veren modeller farklı değerlerde değişkenler denendiğinde kötü sonuçlar vermiştir. Bu sebeple "SubTotal" sütunu çıkarılmıştır. "Model" sütunu kategorik bir sütundur ve aralarında ordinal ilişki olmayan değerler barındırır. Bu sebeple kullanılmak için "One Hot Encoding" işlemi uygulanmıştır. "One Hot Encoding" işleminde her bir değer veri kümesine bir sütun olarak eklenir. "Model" sütununa bu işlem uygulandıktan sonra modellerde denenmiş ve etkisinin az olduğu görülmüştür. Veri kümesine yeni sütunlar eklemek maliyeti arttıran bir işlemdir. Bu sonuçlardan yola çıkarak "Model" sütununun veri kümesinden çıkarılmasına karar verilmiştir. "Model" sütununa uygulanan yöntem benzer ve daha maliyetli sonuçlar verdiği için "SubModel" sütununa da uygulanmıştır. "FuelType" sütunu bir önceki bölümde açıklanan nedenden ötürü silinmiştir. Yukarıda bahsedilen sütunların aksine bahsedilmeyen sütunlar için yine modellerde etkisi denenmiş (Bacward elimination yöntemi ile) ve veri kümesinde kalmasına karar verilmiştir. Satır bazlı olarak hedef değişken olan "GrandTotal" sütunundaki değeri 0'dan küçük yani eksi değerler anlamlı olmayacağından veri kümesinden silinmiştir. Veri kümesinde kullanılan sütunlarda toplamda 540 adet eksik veri bulunmaktadır. Veri kümesindeki toplam gözlem birimi baz alındığında eksik verilerin oranı çok düşük olduğundan ve etkisinin çok az olacağı düşünüldüğünden herhangi bir düzenleme işlemi yapılmamış 540 eksik verinin olduğu satırlar silinmiştir.

2.3.1 Aykırı Değerler (Outliers)

Aykırı değerler, veri setlerinde beklenmeyen veya sıra dışı değerleri temsil ederler ve çoğunlukla diğer veri noktalarından büyük ölçüde uzaktır. Bu uzaklık ciddi farklılıklara sebep olabilmekte ve sonuçları büyük oranda etkileyebilmektedir [17]. Veri analizi ve istatiksel modellemede bu değerlere dikkat edilmeli ve etkisi incelenmelidir.

Bu çalışmada aykırı değerler yakalanmak için 0,05 alt çeyrek ve 0,95 üst çeyrek alınarak eşik değerler hesaplanmış ve aykırı değerler eşik değerlere eşitlenmiştir.

2.3.2 Sütun Dönüşümü (Encoding)

Encoding bilgiyi farklı türlere dönüştürme sürecidir. Modeller kategorik değişkenleri anlayamazlar ve bundan dolayı bu verileri anlayabileceği formatta veri tipine dönüştürmemiz gerekir.

Bu çalışmada "Brand", "Gear" ve "MaintenanceType" sütunlarına One Hot Encoding işlemi uygulanmıştır. Bu işlemde sütundaki değerlerin her biri bir sütun olarak veri kümesine eklenir. Şekil 9' da görülmektedir. Değişkenler birbiri üzerinden üretilebilir olamaması adına bütün One Hot Encoding işlemlerinde drop_first seçeneği true yapılmış ve ilk kolonlar silinmiştir.

Segment	Year	FuelType	Gear	Milage
C	2014	Diesel	Automatic	200130.0
D	2016	Diesel	Automatic	88307.0
C	2017	Diesel	Automatic	221354.0
D	2016	Diesel	Automatic	163614.0
B	2017	Diesel	Automatic	75474.0
C	2016	Diesel	Manual	143909.0

Gear_Half Manual	Gear_Manual	Gear_TRIP TRONIC
0	0	0
0	0	0
0	0	0
0	0	0
0	0	0
0	1	0

Şekil 9: One Hot Encoding işlemi

Veri kümesindeki segment sütununda segmentler arası sıralı farklılıklar olduğundan label encoding işlemi uygulanmıştır.

Model	SubModel	Segment	Year	FuelType	
COROLLA	1.4 D-4D TOUCH M/M	C	2014	Diesel	Au
A4	2.0 TDI QUATTRO 190 HP S TRONIC	D	2016	Diesel	Au
EGEA	COMFORT 1.6 MJET 120HP DCT	C	2017	Diesel	Au
PASSAT	1.6 TDI BMT 120 PS TRENDLINE DSG	D	2016	Diesel	Au
ACCENT	BLUE 1.6 CRDI MODE PLUS OTM	B	2017	Diesel	Au

VendorID	Segment	Year	CityID
929	1	2014	63
7990	2	2016	63
540	1	2017	63
7854	2	2016	63
4390	0	2017	63

Şekil 10: Label Encoding İşlemi

3. MODELLEME

Modelleme, karmaşık problemleri anlamamıza ve geleceği tahmin etmemize yardımcı olan güçlü bir araçtır [18, 19]. Bu aşama proje amacına yönelik eğitim ve tahmin adımlarını içeren bir yöntemdir. Veri kümesindeki veriler kullanılabilir hale getirildikten sonra bu aşamada modellere verilerek eğitim ve tahmin işlemleri yapıldığından proje için önemli bir aşamadır [5]. Bu çalışmada literatür taraması sonucu bulunan maliyet analiz projelerinde kullanılan, benzer regresyon projelerinde iyi sonuçlar vermiş ya da karmaşık verilerde iyi sonuçlar veren çeşitli modeller denenmiştir. Kullanılan modeller: Support Vector Machine, Random Forest, XGBoost, Polynomial Regresssion, LightGBM ve CatBoost modelleridir. Model değerlendirmesi için literatürde yapılan benzer çalışmalarda kullanılan R^2 ve ortalama kare hatası ölçütleri kullanılmıştır.

3.1 Support Vector Machine (SVM)

Özellikle sınıflandırma ve regresyon problemlerini çözmek için kullanılan güçlü bir makine öğrenmesi algoritmasıdır. SVM, çekirdek fonksiyonları sayesinde veriyi yüksek boyutlu uzaylarda ele alabilir [20]. Veri kümemizin karmaşık ve büyük boyutlu olmasından ve bu algoritmanın da benzer problemlerde iyi sonuç verdiği dolayısıyla çalışmamızda kullanılmıştır.

3.2 Random Forest

Yaygın olarak kullanılan ve büyük verilerde başarılı sonuç veren bir topluluk (ensemble) öğrenme modelidir. Model birden çok karar ağacını bir araya getirerek genelleme yeteneği yüksek bir model oluşturur [21]. Literatürde de benzer çalışmada örneği olduğundan veri kümesine uygun görülmüş ve projede kullanılmasına karar verilmiştir.

3.3 XGBoost

Özellikle tablo verileri üzerinde yüksek performanslı tahmin modelleri oluşturmak için kullanılan bir algoritmadır. Regresyon problemlerinde ve karmaşık veri kümelerinde başarılı olduğu görülmektedir [22]. Bu nedenle projede kullanılması uygun bulunmuştur.

3.4 Polynomial Regression

Doğrusal regresyonun genişletilmiş bir biçimidir ve doğrusal olmayan ilişkileri modellemek için kullanılır. Doğrusal regresyonun aksine, bağımsız değişkenlerin mertebeleri kullanılarak daha esnek ve karmaşık bir model elde edilir. Literatürde benzer bir çalışma örneği yoktur ancak farklı türde regresyon problemlerinde karmaşık veri kümelerinde başarılı sonuçlar verebildiği için projede kullanılmasına karar verilmiştir.

3.5 LightGBM

Microsoft tarafından geliştirilen bir açık kaynaklı makine öğrenmesi modelidir. Benzer modellere nazaran daha hızlı ve verimli çalışmaktadır. Elimizdeki veri kümesi üzerinde literatürde gelişmiş ağaç modellerinin başarılı olduğu görüldüğünden deneme amaçlı projede kullanılması uygun görülmüştür [23].

3.6 CatBoost

Kategorik değişkenler üzerinde etkili olan açık kaynaklı bir makine öğrenimi çerçevesidir. Performansı ve özellikle kategorik değişkenler için otomatik işleme özelliği ile öne çıkar [13]. Karmaşık veri kümelerinde de başarılı olması nedeni ile projede kullanılması uygun görülmüştür.

Tablo 2: Model Başarı Değerlendirmeleri

Model	R2 Değeri	Ortalama Kare Hatası
Support Vector Machine	0.920155	30690.149
Random Forest	0.999026	374.172
XGBoost	0.997617	915.827
Polynomial Regression	-615159	2.364
LightGBM	0.999020	376.307
CatBoost	0.999380	237.990

4. DEĞERLENDİRME

Proje kapsamında bir araç kiralama şirketinin bakım maliyet verileri üzerinde ön işleme, analiz ve görselleştirme işlemleri yapılmıştır. Projede iyi sonuçlar alınması amacıyla literatür taraması yapılmıştır. Yapılan araştırma sonucu benzer projelerde makine öğrenmesi modellerinin özellikle ileri seviye ağaç modellerinin etkili olduğu bilgisi edinilmiş regresyon algoritmalarından Support Vector Machine, Random Forest, XGBoost, Polynomial Regression, LightGBM, CatBoost kullanılmasına karar verilmiştir. Benzer çalışmalardaki başarıların karşılaştırılabilmesi açısından literatürdeki çalışmalarda kullanılan R^2 ve ortalama kare hatası metrikleri kullanılmıştır. R^2 değeri, bir regresyon modelinin bağımsız değişkenlerinin bağımlı değişkeni ne kadar iyi açıkladığını ölçen istatistiksel bir başarı ölçüsüdür. Ortalama Kare Hatası, bir regresyon modelinin tahminlerinin gerçek verilerden ne kadar uzak olduğunu ölçen bir istatistiksel ölçüdür. Ortalama kare hatası değeri modele bağlı olarak değişebilir ve bağlamı önemlidir. Bir hata değeri başka bir bağlamda kabul edilebilirken, aynı hata başka bir durumda yüksek olabilir. Literatür taraması sonucu elde edilen model başarısı ve bizim çalışmamızda bulunan sonuçlar aşağıdaki tabloda verilmiştir. İki algoritmada (LightGBM ve Polynomial Regression) benzer bir çalışmada aynı metriklere sahip sonuçlar bulunamamış ancak literatür taraması sonucu farklı yapıdaki benzer veri kümelerinde iyi sonuç verdiği bilgisinden yola çıkarak bizim projemizde de kullanılmış ve sonuçlar paylaşılmıştır.

Çalışmada değinilecek bir başka nokta ise proje gelişim sürecindeki tüm adımlar için gelişmiş yapay zekâ teknolojileri çözümleri de mevcuttur ve bunlardan biri veri kümesi üzerinde uygulanacak, sonuçları alınacak ve sonuçlar bir sonraki bölümde karşılaştırılacaktır. Gelişmiş yapay zekâ teknolojilerine örnek verecek olursak Optiwisdom şirketine ait olan kolay.ai motoru fiyat belirleme, alış ve satış tahmini, talep tahmini gibi konularda yardımcı olmasının yanında müşteri analitiği ile ilgili, gelir ve gider analitiği ve müşteri segmentasyonu yaparak rekabetçi piyasada yardımcı olacak birçok yeteneği barındırmaktadır [24].

Tablo 3: Karşılaştırmalı Model Başarı Sonuçları

Model	Bulunan R^2 Değeri	Literatür R^2 Değeri
Support Vector Machine	0.920	0,923
Random Forest	0.999	0,966
XGBoost	0.997	0,929
Polynomial Regression	-615	-
LightGBM	0.999	-
CatBoost	0.999	0,972

5. SONUÇ

Bu projede bir araç kiralama şirketi verileri kullanılmış, veriler üzerinde önişleme teknikleri uygulanmış ve makine öğrenmesi modelleri kullanılmıştır. Literatür taramasından başlanıp modellerin hazırlanan veri üzerinde denenmesine kadar olan son bölüme kadar adım adım ilerlenen bu süreçte, makine öğrenmesi modelleri ile yapılan maliyet tahmininin ve maliyet analizinin şirketler için önemi ve şirketlere sağlayacağı iyileşmeleri ele almıştır.

Projenin ilk etabında maliyet analizi ve tahmini ile ilgili literatür detaylı bir şekilde incelenmiştir. İncelenen mevcut literatür, şirketlerin sürdürülebilir kaynak yönetimi ve rekabet avantajı elde etmesi için maliyet tahmini ve maliyet tahmini doğrultusunda planlama yapmasının kaçınılmaz olduğunu ortaya çıkarmıştır.

Projede değinilen bir başka husus da literatür taraması sonucu önceden benzer problemler ile ilgili yapılan çalışmalarda uygulanmış makine öğrenmesi modelleri ile bu projede uygulanan model başarı sonuçlarının karşılaştırılmasıdır. Bu karşılaştırma, şirketlerin bir sonraki yıllardaki bütçesini planlarken en doğru sonuçları alabileceği makine öğrenmesi modellerinin seçilmesi için önemli fikirler vermektedir.

Projede veri kümesi ve analizi başlığı altında veri kümesi açıklanmıştır. Veri kümesi üzerinde gerekli görülüp uygulanan veri önişleme teknikleri detaylıca açıklanmıştır. Veri analizi bölümünde "Tableau" uygulaması kullanılarak grafikler çizdirilmiş ve değişkenlerin hedef değişken olan toplam maliyet (GrandTotal) için önemi görsel olarak anlatılmıştır.

Çalışmada odaklanılan bir başka nokta üzerinde işlemler yaptığımız ve çeşitli model başarı sonuçlarını paylaştığımız veri kümesini gelişmiş bir yapay zekâ motoru olan kolay.ai üzerinde denemek ve karşılaştırma yapmaktır. Vurgulanması gereken en önemli hususlardan biri, veri bilimciler tarafından veri kümesi üzerinde manuel olarak programlama dilleri ve algoritmalarla yapılan çalışma sürecinin zorluğu ve zaman maliyetini önemli oranda indirgemesidir. Veri kümesi yapay zekâ motoruna uygun formata çevrilip sisteme yüklendikten daha geniş çaplı analiz ve sonuçları oldukça kısa bir sürede sonuçlar çıkmaktadır. Burada sağlanan en önemli faydalar zaman maliyetini oldukça düşürmesinin yanı sıra daha başarılı sonuçlar verebilmesidir.

Son olarak bu çalışma, maliyet analizi ve makine öğrenmesi modelleri ile maliyet tahminin şirketlere önemli faydalar sağladığını ve doğru yapılan maliyet tahmini çıktılarının şirketlere rekabet avantajı sağlayabileceğini veya rekabet avantajını arttırabileceğini göstermektedir. Yapılan çalışma, şirketlerin bu yapay zekâ teknolojilerini etkili bir biçimde kullanması ile bütçelerinde efektif bir genişlik oluşturmasında yardımcı olacaktır.

6. KAYNAKÇA

[1] Longbing Cao. 2017. Data science: A comprehensive overview. ACM Comput. Surv. 50, 3, Article 43 (June2017), 42 pages http://dx.doi.org/10.1145/3076253

[2] Hamed Jahani, Richa Jain and Dmitry Ivanov. 2023. Data science and big data analytics: a systematic review of methodologies used in the supply chain and logistics research. (May 2023) https://doi.org/10.1007/s10479-023-05390-7

[3] Sanaz T. Hashemi, Omid M. Ebadati and Harleen Kaur. 2020. Cost estimation and prediction in construction projects: a systematicreview on machine learning techniques. SN Applied Sciences (September 2020) 2:1703 https://doi.org/10.1007/s42452-020-03497-1

[4] Eleonora Florian, Fabio Sgarbossa and Ilenia Zennaro. 2021. Machine learning-based predictive maintenance: A cost-oriented model for implementation, Volume 236 (June 2021) https://doi.org/10.1016/j.ijpe.2021.108114

[5] David Donoho. 2017. 50 Years of Data Science. (December 2017) DOI: 10.1080/10618600.2017.1384734

[6] Indla, Raghavendra Kumar. 2021. An Overview on Amazon Rekognition Technology (May 2021). Electronic Theses, Projects, and Dissertations. 1263. https://scholarworks.lib.csusb.edu/etd/1263

[7] Steve New. 2007.Celebrating the enigma: the continuing puzzle of the ToyotaProduction System. International Journal of Production Research,Vol. 45, No. 16, (August 2007) https://doi.org/10.1080/00207540701223386

[8] Wubeshet Woldemariam (2022) A framework for transportation infrastructure cost prediction: a support vector regression approach, Transportation Letters, (October 2022), 14:9, 997-1003, DOI: 10.1080/19427867.2021.1985895

[9] Pai-Hsuen Chen, Chih-Jen Lin and Bernhard Schölkopf. 2005.A tutorial onn-support vector machines, Appl. Stochastic Models Bus. Ind.,(2005) DOI: 10.1002/asmb.537

[10] Haytham H. Elmousalami. 2021. Comparison of Artificial Intelligence Techniques for Project Conceptual Cost Prediction: A Case Study and Comparative Analysis. IEEE TRANSACTIONS ON ENGINEERING MANAGEMENT, VOL. 68, NO. 1, (February 2021) DOI: 10.1109/TEM.2020.2972078

[11] Dima Alberg, Mark Last and Abraham Kandel. 2012. Knowledge discovery in datastreams with regression treemethods. Volume 2, (January/February 2012) DOI: 10.1002/widm.51

[12] Sasanka Katreddi, Arvind Thiruvengadam, Gregory J. Thompson and Natalia A. Schmid. 2023. Mixed Effects Random Forest Model for Maintenance Cost Estimation in Heavy-Duty Vehicles Using Diesel and Alternative Fuels, VOLUME 11, (July 2023). DOI: 10.1109/ACCESS.2023.3290994

[13] Muhammad Kamran. 2021. A State of the art Catboost-Based T-Distributed Stochastic Neighbor Embedding Technique to Predict Back-break at Dewan Cement Limestone Quarry. (September 2021). DOI:10.22044/jme.2021.11222.2104

[14] Korcan Doğan, Sacit Arslantekin. 2016. Büyük veri: Önemi, Yapısı ve Günümüzdeki durum. (Haziran 2016) DOI: 10.1501/Dtcfder_0000001461

[15] Şadi Evren ŞEKER, "Fleet Maintenance Dataset", Mayıs 2023, 10.34740/kaggle/ds/3067824

[16] Salvador García* , Sergio Ramírez-Gallego, Julián Luengo, José Manuel Benítez and Francisco Herrera. 2016. Big data preprocessing: methods and prospects (November 2016). DOI: 10.1186/s41044-016-0014-0

[17] Cheng Fan, Meiling Chen, Xinghua Wang, Jiayuan Wang and Bufu Huang. 2021. A Review on Data Preprocessing Techniques Toward Efficient and Reliable Knowledge Discovery From Building Operational Data (March 2021). DOI: 10.3389/fenrg.2021.652801

[18] A. Salmi, P. David, E. Blanco ve J.D. Summers. 2016. A review of cost estimation models for determining assemblyautomation level. (June 2016) . http://dx.doi.org/10.1016/j.cie.2016.06.007

[19] Frank Bodendorf , Philipp Merkl, Jorg Franke. 2021. Intelligent cost estimation by machine learning in supply management: Astructured literature review. (August 2021). https://doi.org/10.1016/j.cie.2021.107601

[20] Marti A. Hearst, S. T. Dumais ve arkadaşları. 1998. Support vector machines. (July-August 1998). DOI: 10.1109/5254.708428

[21] Steven J. Rigatti ve arkadaşları. 2016. Random Forest. (November 2016). https://doi.org/10.17849/insm-47-01-31-39.1

[22] Adeola Ogunleye , Qing-Guo Wang . 2020. XGBoost Model for Chronic Kidney Disease Diagnosis . (Nov.-Dec. 2020). DOI: 10.1109/TCBB.2019.2911071

[23] Amir Izadi, Mohammad Nabipour ve Omid Titidezh. 2019. Cost Models and Cost Factors of Road Freight Transportation:A Literature Review and Model Structure. (October 2019). DOI: 10.1080/16168658.2019.1688956

[24] Şadi Evren ŞEKER. 2023. KOBİ'ler için Yapay Zeka Çözümü: Kolay.AI, YBS Ansiklopedi, v.11, is 1, pp. 24 - 46, 2023