Uludağ Üniversitesi Mühendislik Fakültesi Dergisi, Cilt 31, Sayı 1, 2026 ARAŞTIRMA

DOI: 10.17482/uumfd.1733355

AKARYAKIT SEKTÖRÜNDE İŞLEM ANORMALLİKLERİNİN VERİ MADENCİLİĞİ YÖNTEMLERİ İLE SINIFLANDIRILMASI

Sabahattin Mert BERKMEN [*]
Sabahattin Kerem AYTULUN [**]

Alınma: 02.07.2025; düzeltme: 21.02.2026; kabul: 08.03.2026

Öz: Akaryakıt sektöründe veri madenciliği uygulamaları her geçen gün gelişmekte ve yaygınlaşmaktadır. Sektörde kullanılan yöntemler ve yapılan çeşitli analizler sayesinde, akaryakıt hırsızlığı, operasyonel anormallikler, dolum sırasında meydana gelen miktar aşımları ve aşırı dolum sonrası yaşanan taşma gibi kritik konular izlenerek gerekli aksiyonlar alınmaktadır. Bu çalışmada, bir petrol şirketinin verileri kullanılarak daha önceden belirlenmiş dört önemli kritik kategori için sınıflandırma çalışması gerçekleştirilmiştir. Veri setine ön işleme uygulanmış, analize katkısı olmayan değişkenler veri setinden çıkarılmış ve eksik veriler tamamlanarak analiz için uygun bir hale getirilmiştir. Uygulama aşamasında RAPIDMINER (v.9.10) yazılımından yararlanılmıştır. Veri madenciliği sınıflandırma yöntemlerinden k-en yakın komşu algoritması, Rastgele Orman Algoritması, Gradient Boosted Algoritması, ADABOOST Algoritması ve Karar Ağacı (J48) Algoritması kullanılarak sınıflandırma işlemleri gerçekleştirilmiş ve modellerin başarısı çeşitli ölçütler kullanılarak değerlendirilmiştir.

Anahtar Kelimeler: Veri Madenciliği, Sınıflandırma, Akaryakıt Sektörü, İşlem Anormallikleri

Classification Of Process Abnormalities in the Fuel Industry by Data Mining

Abstract: Data mining applications in the fuel industry are advancing and becoming more widespread with day by day. Through the methods and various analyses used in the sector, critical issues such as fuel theft, leaks, quantity overruns during refueling, and overflow after overfilling are monitored, and necessary actions are taken. In this study, classification analysis was conducted for four predefined critical categories using the data of a petroleum company. Preprocessing was applied to the dataset, irrelevant variables that did not contribute to the analysis were excluded, and missing data were completed to make the dataset ready for analysis. The RAPIDMINER (v.9.10) software was utilized during the implementation phase. Classification methods in data mining, including the k-nearest neighbors algorithm, Random Forest Algorithm, Gradient Boosted Algorithm, ADABOOST Algorithm, and Decision Tree (J48) Algorithm, were employed to perform classification. The performance of the models was evaluated using various success metrics.

Keywords: Data Mining, Classification, Fuel Industry, Process Abnormalities

1. GİRİŞ

Günümüz dünyasında, gelişen teknolojiler ile birlikte işletmeler ham verileri veri tabanlarında tutmaktadırlar. Bu veriler gün geçtikçe artmakta olup, işletmelerin daha kaliteli hizmet verebilmesini sağlayan başlıca unsurlara dönüşmektedir. Veri setleri, ilk halleri göz önüne alındığında birden çok sorun barındırabilmektedir. Veri madenciliği algoritmalarının

[*] Sabahattin Mert Berkmen İstanbul Beykent Üniversitesi, Mühendislik Mimarlık Fakültesi, Endüstri Mühendisliği Bölümü ,34396, İstanbul
[**] Sabahattin Kerem Aytulun İstanbul Arel Üniversitesi, Mühendislik Fakültesi, Endüstri Mühendisliği Bölümü, 34537, İstanbul

İletişim Yazarı: Sabahattin Mert Berkmen (mertberkmen@gmail.com)

performansını artırmak ve etkili sonuçlar alabilmek için bu veri setlerinin ön işleme süreçleri ile birlikte analiz edilmesi ve uygun hale getirilmesi gerekir. Büyük veri döneminin başlamasıyla birlikte kullanıma sunulan yeni teknolojiler sayesinde büyük veriler ile çalışma fırsatı elde edilmiştir. Bu durum işletmelere, sınıflandırma, kümeleme vb. önemli analizler yapma imkanı da yaratmış, iş süreçleri ve karar alma mekanizmaları da bu gelişme paralelinde değişime uğramıştır. Bu sektörlerden biri de enerji ihtiyacının büyük bir bölümünü teşkil eden akaryakıt sektörüdür. Akaryakıt sektöründe veriler otomasyon sistemleri üzerinden veri tabanlarına aktarılmaktadır. Akaryakıt sektöründe, veriler işlenip Enerji Piyasası Düzenleme Kurumu (EPDK) kuralları göz önünde bulundurularak, çevre ve maliyet açısından kritik öneme sahip olan operasyonel anormallikler gibi unsurlar ile diğer önemli kategoriler belirlenmekte ve bu kategoriler bazında sınıflandırma yapılmaktadır. Veri madenciliği 1990 yıllında ortaya çıkmış bir kavramdır. Ancak 2000 yıllından itibaren çeşitli sektörlerde kullanıldığı görülmüştür. Çeşitli sektörlerle ilgili çalışmalar kronolojik olarak aşağıda açıklanmıştır. Boland ve arkadaşları ABD ordusu Mühendisler Birliği'ndeki su kullanımını veri madenciliği yöntemleri ile tahminlenerek açıklamışlar ve bu konu hakkında tavsiyelerde bulunmuşlardır (Boland ve diğ., 1981). Sforna, enerji alanında faaliyet gösteren bir işletmenin müşteri faturalandırma işlemlerini desteklemek amacıyla oluşturduğu büyük ve karmaşık veri tabanını kullanarak, veri madenciliği yöntemleriyle gizli bilgileri ortaya çıkarmayı ve keşfedilmemiş kuralları belirlemeyi hedeflemiştir (Sforna, 2000). Özekes, çalışmasında veri madenciliği kapsamında işlevlerine göre farklılık gösteren birliktelik kuralları, sınıflama, regresyon ve kümeleme kavramlarını detaylı bir şekilde açıklamış ve bu kavramların uygulama alanlarına değinmiştir (Özekes, 2003). Chen ve arkadaşları, çalışmalarında tedarik zinciri yöntemlerini ele almış ve paralel koridor düzenine sahip bir dağıtım merkezindeki sipariş toplama problemini çözmek için veri madenciliği yöntemlerini kullanmışlardır. Uygulama kapsamında, dağıtıma çıkacak ürün gruplarını kümeleyerek etkili bir çözüm sunmuşlardır (Chen ve diğ, 2005). Kuşaksızoğlu, tez çalışmasında bir telekomünikasyon şirketinde normal kullanıcılarla sahtekarlık yapan kullanıcıları ayırt etmek amacıyla bir model geliştirmiş ve bu süreçte abonelerin konuşma detayları ile demografik özelliklerini incelemiştir (Kuşaksızoğlu, 2006). Dönmez, dayanaklı tüketim sektöründe faaliyet gösteren bir işletmede var olan birbirlerine benzer bayilerin veri madenciliği yöntemleriyle kümelemiştir. Kümeleme sonucunda ortaya bayilerin pazar stratejileri belirlenerek cirolarının artırılması amaçlanmıştır (Dönmez,2008). Dudas ve arkadaşları yapmış oldukları çalışmada, işletmelerin herhangi bir konuda stratejik karar verme aşamasında veri madenciliği yöntemlerinden ve optimizasyon yöntemlerinin ortaklaşa kullanımından daha iyi sonuçlar elde edebileceğini ortaya koymuşlardır (Dudas ve diğ., 2013). Dominic ve arkadaşları veri madenciliği yöntemlerini kullanarak inşaat maliyetlerinin tahminlenmesi üzerine model geliştirmişlerdir. Toplam 1600 adet inşaat projesi tahmin edilmiş ve modelin başarılı olduğu görülmüştür (Dominic ve Dagbui, 2014). Dalman, akaryakıt sektöründeki en önemli problemlerden birinin sızıntı tespiti olduğunu belirtmiştir. Ancak sızıntı tespiti, tanklarda oluşan kalibrasyon kaynaklı farklarla sıklıkla karıştırılmaktadır. Çalışmada veri madenciliği yöntemlerini kullanarak sızıntı tespitini kolaylaştırmak ve iyileştirmek amaçlanmıştır (Dalman, 2017). Dos Santos ve arkadaşları, çalışmalarında 2009-2018 yılları arasında çeşitli bilimsel veri tabanlarından elde edilen bilimsel çalışmaları incelemiş ve bunları kullanılan veri tabanlarına, tekniklere ve programlama dillerine göre sınıflandırmıştır. Elde edilen sonuçlar ayrıntılı bir şekilde sunulmuştur (Dos santos ve diğ., 2019). Schuh ve arkadaşları yapmış oldukları çalışmada üretim sektöründe süreç planlamada veri madenciliği yöntemlerinin kullanılmasının öneminden bahsetmişlerdir (Schuh ve diğ., 2020). Bu çalışma altı bölümden oluşmaktadır. İlk bölümde araştırmanın kapsamı ortaya konularak konuya ilişkin genel çerçeve sunulmuş ve literatürde yer alan çalışmalar incelenmiştir. İkinci bölümde araştırmada kullanılan veri seti, veri ön işleme süreci ve yöntemsel yaklaşım detaylı biçimde açıklanmıştır. Üçüncü bölümde çalışmada kullanılan sınıflandırma algoritmaları tanıtılmış ve teorik altyapıları ele alınmıştır. Dördüncü bölümde model performanslarının değerlendirilmesinde kullanılan ölçütler açıklanmıştır. Beşinci bölümde elde edilen bulgular sunulmuş ve modellerin

karşılaştırmalı analizleri yapılmıştır. Son bölümde ise araştırma sonuçları tartışılarak genel değerlendirmeler yapılmış ve gelecekte yapılabilecek çalışmalara yönelik öneriler sunulmuştur.

2. MATERYAL VE METOT

Çalışmada kullanılan veriler, Türkiye'de bulunan tüm petrol şirketlerine veri analizi ve teknik anlamda destek sağlayan otomasyon firmasının yazılı onayı doğrultusunda veri tabanlarından başvuru süreci tamamlanarak elde edilmiştir. Bu çalışmada kullanılan veri seti, ilgili petrol ve otomasyon şirketi tarafından temin edilmiş olup, gizlilik sözleşmesi çerçevesinde sağlanmıştır. Veri setinde yer alan istasyon kimlikleri, coğrafi konum bilgileri ve şirket içi özel bilgiler anonimleştirilmiş, doğrudan şirketi veya bir istasyonu tanımlamaya olanak sağlayacak tüm veriler çıkarılmıştır. Anonimleştirme süreci kapsamında, veriler rastgele kodlanarak kimliksiz hale getirilmiş ve analizler bu anonim veri kümesi üzerinden gerçekleştirilmiştir. Ayrıca, veri seti yalnızca yetkilendirilmiş araştırmacılar tarafından güvenli ortamlarda erişilebilecek şekilde korunmuş, veri güvenliğini sağlamak amacıyla dışa veri aktarımı veya paylaşımı yapılmamıştır. Elde edilen veriler bir petrol şirketinin 2022-2023 yılları arasındaki verileri kapsamaktadır. Çalışmada ilk olarak uzmanlar tarafından günlük olarak alarm üreten veri grupları analiz edilmiş, analizler sonucunda kategori bazında sınıflandırılmıştır. Sınıflandırma sonucu elde edilen sonuçlar bir sonraki aşamada kurulacak olan sınıflandırma modelinde eğitim verisi olarak kullanılacaktır. Veri gruplarının sınıflandırılacağı yaklaşık olarak 50 ana kategori bulunmaktadır. 50 kategori içerisinden, firmalar için önem arz eden dört kategori seçilmiş ve çalışma bu dört kategori üzerinden sürdürülmüştür. Bahsedilen kategorilerin neler olduğu ve neleri ifade ettiği bulgular bölümünde Tablo.2'de "Kategori Açıklaması" başlığı altında detaylıca verilmiştir. Bu çalışma, sürekli ve nicel olarak izlenebilen operasyonel anormalliklere odaklanmaktadır. Akaryakıt sektöründe önemli bir problem olmakla birlikte, sızıntılar ve akaryakıt kaçakçılığı çoğunlukla anlık, olay bazlı ve manuel tespit gerektiren durumlar olup, çalışmada kullanılan otomatik ölçüm ve sensör verileriyle düzenli olarak izlenememektedir. Bu nedenle sızıntı kategorisi, veri sürekliliği ve yöntemsel uyumluluk kriterleri göz önünde bulundurularak çalışma kapsamı dışında bırakılmıştır. Çalışmada, söz konusu dört kategori için veri setleri kullanılarak farklı sınıflandırma algoritmaları eğitilmiş ve performansları karşılaştırılmıştır.

Çalışmada analiz aracı olarak RAPIDMINER (v.9.10) programı kullanılmıştır. İşletmenin veri tabanından çekilen veri seti 15 sütun ve 6987 satırdan oluşmaktadır. Bahsedilen veri setine ait özelliklere kısaca değinmek gerekirse; "Sıra No" (Row No), veri setindeki verileri belirli bir düzende numaralandırmak için kullanılmıştır. "Tarih" (Date) verinin veri tabanından alındığı tarihi belirtmektedir. "İstasyon Kodu "(Station Code) ve "İstasyon Adı" (Station Name) petrol şirketlerine özgü olup istasyon adı ve istasyon kodlarını belirtmektedir. Bu veriler çalışmada kullanılan verilerin gizliliği sebebiyle kodlanmıştır. Çalışmamızda İstasyon Kodu (Station Code) ve İstasyon İsmi (Station Name) sütunları analize etki etmeyeceğinden dolayı analizde kullanılmamıştır. "Tank No" istasyonlarda bulunan tank numaralarını belirtmektedir. "Oil Type" akaryakıt türünü, "açılış"(Open) ilgili tarihte tankta satış yapılmadan var olan akaryakıtın brüt miktarını, "kapanış" (Close) ise gün sonu satışlarla beraber tankta kalan brüt miktarı litre cinsinden ifade etmektedir. "Yakıt İkmali" (Refuel) istasyonun ilgili tankına o gün dolum yapılıp yapılmadığını yapıldıysa brüt miktarını litre cinsinden ifade etmektedir. "Satış" (Sell) gün içinde ilgili tanktan yapılan toplam akaryakıt miktarını ifade ederken, "Reductıon" satış miktarına bağlı olarak tanktan azalma miktarını ifade eder. "Fark" (Difference) sütunu ise satış miktarı ile azalma miktarı arasındaki farktır Fark negatif olabileceği gibi pozitif de olabilmektedir. "Rate" ((satış miktarı- fark)/(satış miktarı))*100 formülü ile hesaplanırken satış miktarının farka yüzdesel oranını ifade etmektedir. "SEL" (satış miktarı*0,01)+(16,40) formülü ile hesaplanır. Eğer fark değerimiz SEL değerinden büyük olursa veri tabanında alarm üretir ve o istasyonun detaylı olarak incelenmesi gereklidir. İncelenen veri setleri 4 ana grup altında analizci tarafından sınıflandırılmaktadır.

Tablo 1. Veri Setinin Sınıf Dağılımı ve Örneklem Sayları

Kategoriler	Örnek Sayısı	Yüzde(%)
Sıcaklık Değişimi	4891	20%
Arıza	4891	20%
Tank Kalibrasyonu	7886	32%
Dolum-Tesisat	6987	28%
Toplam	24655	100%

Çalışmada kullanılan veri setinin kategorik dağılımı Tablo.1'de sunulmuştur. Toplam 24.655 örneklemden oluşan set, sınıflar arası dengeli bir dağılım sergilemektedir.

Tank No	Oil Type	Open	Close	Refuel	Reduction	Sell	Difference	Rate (%)	SEL
4	MOTORIN	4171078	3947128	0	223.950	0	-223.950	0	16.400
3	MOTORIN	904907	904940	0	-0.033	226.550	226.583	100.014	18.665
2	MOTORIN	6009908	5599440	0	410.468	0	-410.468	0	16.400
1	MOTORIN	3095417	2914510	0	181.907	590.980	409.073	69.219	22.310
4	MOTORIN	3947128	3819139	0	127.990	0	-127.990	0	16.400
3	MOTORIN	904940	904788	0	0.152	128.480	128.328	99.882	17.685
1	MOTORIN	1543716	1637918	1	252.798	502.710	249.912	49.713	21.427
4	MOTORIN	5218044	9919353	1	155.233	0	-155.233	0	16.400
2	MOTORIN	566351	7987095	1	703.522	0	-703.522	0	16.400
1	MOTORIN	9959393	28095544	1	168.917	851.530	682.613	80.153	24.915
5	MOTORIN	3471953	12794021	1	570.450	757.480	187.030	24.691	23.975
3	MOTORIN	2895480	8554338	1	10.068.470	10.214.710	146.240	1.432	118.547
1	MOTORIN	2974728	15751908	1	12.934.829	12.369.110	334.281	2.519	14.091

Şekil 1:

Veri Deseni Örneği

Veri madenciliği modellerinde eksik verinin temizlenme işlemi "null" olarak adlandırılan eksik veri satırına sahip olan kısmın analizden tamamen çıkarılmasını veya çeşitli yöntemlerle eksik verinin doldurulması sürecini içerir (Doğan ve diğ.,2021). Çalışmamızda kullanılan veri setinde yeterli sayıda gözlem bulunduğundan, eksik veri problemi sınırlı düzeyde kalmıştır. Eksik değer içeren gözlemler, analiz sonuçlarını yanıltabilecek ölçüde olmadıkları için uygun şekilde ele alınmıştır. Bu kapsamda, eksik veri içeren az sayıdaki gözlem, analizde kullanılmayacak şekilde dışlanmış veya uygun yöntemlerle (örneğin ortalama/medyan ile tamamlama gibi) düzenlenmiştir. Veri önişleme sürecinde ayrıca; kategorik değişkenlerin uygun biçimde kodlanması, sayısal değişkenlerin ölçeklenmesi ve aykırı değerlerin tespiti gibi işlemler de gerçekleştirilmiştir. Ayrıca Aykırı değerler görselleştirme ve istatistiksel yöntemlerle incelenmiş, gerektiğinde analiz dışı bırakılmıştır. Tüm bu önişleme adımları, veri kalitesini artırmak ve modelleme sürecinde daha sağlıklı sonuçlar elde etmek amacıyla uygulanmıştır. Çalışmamızda toplam 6987 adet veri RM "retrieve" sekmesi aracılığı ile aktarılmıştır. Aktarılan veride birden fazla öznitelik bulunmakta olup "set_role" özelliği ile hangi öz niteliğe göre sınıflandırılacağı belirlenmiştir. Çalışmamızda bu özellik kategori olarak belirlenmiştir. Özellik belirlendikten sonra çalışmada veri setinin istenilen sayıda alt kümlere bölünmesi için "split_data" özelliği kullanılmıştır. "split_data" operatörü, bir veri setini girdi olarak alır ve veri setinin alt kümelerini

çıkış portları aracılığıyla iletir. Alt kümelerin (veya bölümlerin) sayısı ve her bölümün göreli boyutu, "partitions" parametresi aracılığıyla belirtilir. Tüm bölümlerin oranının toplamı 1 olmalıdır. Çalışmada, eğitim ve test verileri, ana veri setinin, genel kabule uygun olarak sırasıyla %70 ve %30 olarak ayrılmasıyla elde edilmiştir. "split_data" özeliğinden sonra veri içerisinde yer alan eksik satıra sahip veri satırları "filter_examples" aracılığıyla analizimizden çıkarılmıştır. "Select_attributes" bir örnek kümenin alt kümlerini seçerek diğer niteliklerin kaldırılmasını sağlayan analiz aracıdır. Çalışmamızda veri seti içerisinde yer alan "date", "station_code" ve "station_name" nitelikleri analizimizi etkilemediğinden ve gereksiz yer kaplamaması adına "select_ attributes" aracı kullanılarak çıkarılmıştır. Veri önişleme adımları tamamlandıktan sonra sınıflandırma modelinin belirlenmesi aşamasına geçilmiştir. Sınıflandırma modellerinin istatistiksel başarısını ölçmek amacıyla "cross_ validation" tercih edilmiştir. "Cross_validation" operatörü, bir sınıflandırma modelinin istatistiksel performansını tahmin etmek için çapraz doğrulama gerçekleştirir ve belirli bir sınıflandırma algoritması tarafından öğrenilen bir modelin ne kadar doğru performans göstereceğini tahmin etmek için kullanılır. Eğitim alt süreci ve test alt süreci olmak üzere iki alt süreci vardır. Eğitim alt süreci, bir modeli eğitmek için kullanılır.

3. KULLANILAN ALGORİTMALAR

k-Nearest Neighbors *(k*-NN) algoritması, küçük ve düşük boyutlu etiketlenmiş veri setlerinde basitlik ve yorumlanabilirlik açısından avantajlı bir sınıflandırma algoritmasıdır. Algoritma, bir veri noktasının sınıfını belirlemek için en yakın *k* komşu noktanın etiketlerini baz alarak karar verir ve bu noktalar arasındaki benzerlikleri mesafe ölçümleriyle değerlendirir (Şimşek, 2019).

Karar Ağacı algoritmaları diğer sınıflandırma yöntemleri ile karşılaştırıldığında anlaşılması kolay ve başarı oranı yüksek olan sınıflandırma algoritmasıdır (Agrawal ve diğ., 1993). Algoritmanın avantajları olduğu kadar dezavantajları da bulunmaktadır. Dezavantajlara bakıldığında karar ağacında analiz edilmesi istenen veri grubu ne kadar fazla ise dallanma sayısı da o kadar fazla olacağından analiz süreleri uzamakta ve maliyetler artmaktadır (Demirel ve Yakut, 2019). Yapısına bakıldığında tipik bir ağacın yapısına benzemekte olup dal, yaprak ve karar düğümlerinden oluşmaktadır. İlk adım olarak algoritma karar düğümleri ile başlar bu adımda veri setlerine kazanılan bilgiyi veya entropi temelli çeşitli ölçütler kullanılır. Bu adımdan çıkan sonuçlar doğrultusunda dallar elde edilir. Sınıflandırılmanın sonuçlandırılması için elde edilen her dalın yaprağa ulaşması gereklidir. Karar ağaçlarında karar düğümü adımı sonrasında sınıflandırılacak olan verinin hangi özelliğini kullanarak dallandırılacağını belirlemek için Bilgi Kazanımı (Information Gain), Kazanım Oranı (Gain Ratio), Gini İndeksi (Gini Index) ölçütlerinden yararlanılmaktadır.

Karar ağacı (J48) algoritması C 4.5 algoritması olarak da bilinmektedir. İlk defa 1983 yılında Ross Quinlan tarafından önerilmiş bir karar ağacı algoritmasıdır. Algoritma sınıflandırma aşamasında kayıp değer olan verileri sınıflandırma sürecine dahil etmeyerek sadece var olan veri üzerinden sınıflandırma sürecini devam ettirmektedir. Algoritmanın kategorik değişkenlerin sınıflandırılmasında oldukça iyi sonuçlar verdiği görülmüştür. Sınıflandırılacak olan verinin hangi özelliğinin kullanarak dallandırılacağının belirlenmesi için özellik seçim ölçütlerinden bilgi kazanımından (Information Gain) yararlanır, ancak herhangi bir özelliğin değer çeşidinin olması, kriteri, bu özelliğin kullanılarak dallanmaya zorlar. Bu da bir taraflılık oluşturur. Bu sebeple ölçek olarak GINI indeksinin kullanımı yaygınlaşmıştır. Rastgele orman algoritması, diğer karar ağacı algoritmalardan farklı olarak sınıflandırma süreci içerisinde birden fazla karar ağacı oluşturarak sınıflandırma yapan denetimli bir öğrenme algoritmasıdır. Bir düğümün en iyi sonucunu bulup parçalara ayırıp dallar elde etmek yerine rastgele seçilen bir düğümün alt kümlerinden en iyi özelliği arayan algoritmadır. Algoritmanın yaygın olarak kullanılmasının nedeni arasında karmaşık düzensiz büyük veri grupları ve kategorik veri gruplarının sınıflandırılmasında oldukça iyi sonuçlar elde etmesidir (Aydın, 2019).

Gradient Boosted Trees algoritması, rastgele orman algoritmasında olduğu gibi karar ağacı temelli bir algoritmadır. Algoritmanın rastgele orman algoritmasından tek farkı son sınıflandırma kararının rastgele orman algoritmasında olduğu gibi bir düğüme bağlı kalmaması, tüm düğümlerin ve ağaçların sınıflandırma kararında etkin olarak son adımda kontrol edilmesidir.

ADABOOST algoritması, Freund ve Schapire tarafından 1996 yılında önerilmiştir (Vezhnevets ve Vezhneets, 2005) tarafından her karar ağacı turunda önceki turda elde edilen modelin başarısına bağlı kalmadan her turda yeni model eğitime dahil edilir. Ayrıca her karar ağacı turunda yanlış olarak sınıflandırılan veriler eğitim amacıyla tekrardan toplanır. Burada amaç bir sonraki karar turunda oluşturulacak olan modele geri besleme yolu ile fayda sağlamaktır. Buradaki fikir, sonraki modellerin önceki modellerin yaptığı hataları telafi edebilmesidir.

4. PERFORMANS ÖLÇÜTLERİ

Sınıflandırılan modelin hangilerinin diğer modellere göre iyi sonuçlar verdiğini belirlemek amacıyla performans ölçütleri kullanılmaktadır. Bu ölçütlerden en önemlileri doğruluk hata oranı, duyarlılık oranı, kesinlik oranı ve F ölçütüdür.

4.1. Karmaşıklık Matrisi

Karmaşıklık matrisi, sınıflandırma algoritmalarının başarısının değerlendirilmesi için kullanılan performans değerlendirme aracıdır. Tablo.2'de karmaşıklık matrisi ile ilgili bir gösterim yapılmıştır.

Tablo 2. Karmaşıklık Matrisi (Confusion Matrix)

		Tahmin Edilen Sınıf	
		Pozitif	**Negatif**
Gerçek Sınıf	**Pozitif**	(TP)	(FP)
	Negatif	(FN)	(TN)

TP: True Positives (Doğru Pozitif), TN: True Negatives (Doğru Negatif), FP: False Positives (Yanlış Pozitif), FN: False Negatives (Yanlış Negatif)

4.2. Kesinlik

Kesinlik oranı literatüre bakıldığında tamlık olarak da bilinmektedir. Algoritmanın pozitif olarak tahmin ettiği değerlerin gerçekten pozitif olduğunu ölçmeye yarayan bir yöntem olarak karşımıza çıkmaktadır [18]. Kesinlik değeri eşitlik-1 de görüldüğü üzere sınıflandırıcı tarafından pozitif olarak sınıflandırılan demet sayısının (TP), sınıflandırıcı tarafından sınıflandırılmış tüm pozitif demetlere bölünmesiyle bulunur.

$$\text{Kesinlik} = \frac{TP}{TP + FP} \tag{1}$$

4.3. Duyarlılık

Gerçek sınıf içerisinde bulunan pozitif demet olarak nitelendirilmiş olan değerlerin sınıflandırıcı tarafından ne kadarının gerçekten pozitif olduğunu belirlemeyi amaçlayan performans ölçütüdür. Eşitlik-2 de görüleceği üzere true positive değerlerinin true positive ve false negative toplamlarına bölünmesiyle elde edilir.

$$\text{Duyarlılık} = \frac{TP}{TP + FN} \tag{2}$$

4.4. Doğruluk

Doğruluk-hata oranı diğer yöntemlere göre modelin başarısı için en çok kullanılan ve tercih edilen yöntemdir. Eşitlik-3'de görüleceği üzere bütün veriler içerisinde doğru olarak sınıflandırılan veri oranını gösterir.

$$\text{Doğruluk} = \frac{TP + TN}{TP + FP + TN + FN} \tag{3}$$

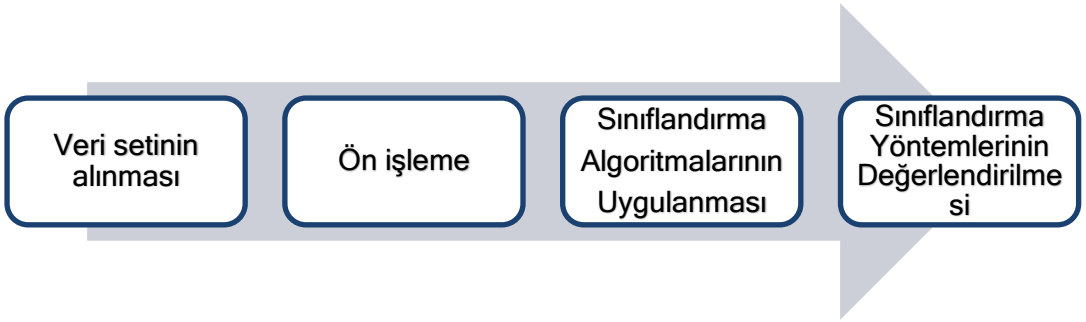
4.5. F1 Skoru

Veri madenciliği performans ölçütü yöntemlerinde bazı modeller için kesinlik ve duyarlılık ölçümlerini ayrı ayrı olarak yorumlamak modelin performansı hakkında yeterince bilgi vermemektedir. Bu yüzden her iki yöntemin aynı anda değerlendirilmesi için F-ölçütü kavramı geliştirilmiştir. F- ölçütü kesinlik ve duyarlılık ölçütlerinin harmonik ortalaması olarak tanımlanmaktadır (Coşkun ve Baykal, 2011).

$$\text{F1 Skoru} = \frac{2 * \text{Duyarlılık} * \text{Kesinlik}}{\text{Duyarlılık} + \text{Kesinlik}} \tag{4}$$

5. ARAŞTIRMA BULGULARI

Bu çalışmada Türkiye'de petrol şirketlerine veri analizi desteği ve teknik destek hizmeti sağlayan, otomasyon firmasının veri tabanında bulunan ve Türkiye'de faaliyet gösteren bir petrol şirketinin istasyonlarına ait olan otomasyon verileri kullanılarak analizler yapılmıştır. Ön işleme sürecinden geçirilen veriler Şekil.2 de gösterilen adımlar doğrultusunda yapılmıştır.



Şekil 2:
Uygulama Süreç Diyagramı

İstasyondaki akaryakıt tanklarının anlık verileri, kurulan cihazlar sayesinde merkeze ayrı ayrı iletilmektedir. Aktarılan veri içerisinde istasyonda tanklarda meydana gelen sorun (kalibrasyon, stok aşımı, dolum, arıza vs.) doğrultusunda alarm üretilmekte olup bu alarmın incelenmesi gerekmektedir. Alarmların aynı gün içerisinde uzmanlar tarafından hatasız ve eksiksiz şekilde incelenmesi ve yaklaşık olarak sorunun kaynağına bağlı olacak şekilde 50 kategori arasından seçilecek en uygun kategoriye atanması gerekmektedir. İncelenen veriler daha sonra Enerji Piyasası Düzenleme Kurumu (EPDK)'ya iletilmektedir. İletilen her yanlış analiz sonucunda şirket EPDK tarafından cezalandırılabilmektedir. Çalışmamızda işletme için önemli

ve iş yükü olarak analizlerin büyük bir bölümünü oluşturan 4 önemli kategori (tank kalibrasyonu, dolum-tesisat, sıcaklık değişimi, arıza) seçilmiştir.

Tablo 3. Kategori Açıklama

Kategori	Açıklama
Sıcaklık Değişimi	Tank içerisinde bulunan yakıtın mevsimsel nedenlerle veya tank içerisine dökülen yakıt ile etkileşime girmesi sonucunda sıcaklık farklılıkları sonucunda farklar oluşmakta olup bu kategori sıcaklık değişimi kategorisi olarak adlandırılmaktadır. Tankta bulunan yakıtın sıcaklığı arttıkça envanter artışı sıcaklık düştükçe envanter azalışı görülmektedir. Manipülasyon çıkarımlarında sıcaklık değişimi kategorisi önemli bir yer tutmaktadır.
Tank Kalibrasyonu	Tank kalibrasyonu, tank içerisinde bulunan yakıt hacminin belirli bir referans değerine göre değerlendirilmesi işlemi olarak adlandırılmaktadır.
Dolum-Tesisat	İstasyonlara tank bazlı olarak yapılan dolumlar sonrası oluşan alarm kategorisidir.
Arıza	Arıza kategorisi istasyonlarda meydana gelen arızaların tespiti sonrasında farklar yaratan alarm kategorisidir. Çalışmamızda arıza kategorisi ATG, probe, router ve elektrik arızaları olmak üzere 4 alt grup altında analizlere dahil edilmiştir.

EPDK'ya iletilen sonuçlarda hatanın en aza indirilmesi amacıyla uzmanlar tarafından yapılan analizlerin sonuçları kullanılarak seçilen 4 kategori için ayrı ayrı veri madenciliği sınıflandırma yöntemleri ile uzman kararlarının doğruluğunun ve model başarısının ölçülmesi amaçlanmıştır. Sınıflandırma algoritması olarak k-En Yakın Komşu Algoritması, Rastgele Orman Algoritması, Gradient Boosted Treess Algoritması, ADABOOST Algoritması ve Karar Ağacı (J48) Algoritması veriler üzerine uygulanmıştır. Gerekli veri madenciliği için uygun hale getirilen veri %30 test veri seti %70 eğitim veri seti olarak ayrılmıştır. Oluşturulan modellerin performans ölçümleri gerçekleştirilmiştir. Kullanılan algoritmalara ait karmaşıklık matrisleri Tablo'4'de verilmiştir.

Tablo 4. *k*-NN Algoritması Karmaşıklık Matrisi

***k*-NN Algoritması Kalibrasyon Modeli Karmaşıklık Matrisi**			***k* -NN Algoritması Dolum-Tesisat Modeli Karmaşıklık Matrisi**		
positive_predictive_value: 67.79% +/- 5.08% (micro average: 67.58%) (positive class: 1)			positive_predictive_value: 80.18% +/- 1.88% (micro average: 80.15%) (positive class: 1)		
ConfusionMatrix:			**ConfusionMatrix:**		
True	**0**	**1**	**True**	**0**	**1**
0	3259	503	**0**	1782	258
1	366	763	**1**	566	2285
***k* -NN Algoritması Sıcaklık Değişimi Modeli Karmaşıklık Matrisi**			***k* -NN Algoritması Arıza Modeli Karmaşıklık Matrisi**		
positive_predictive_value: 61.84% +/- 4.95% (micro average: 61.79%) (positive class: 1)			positive_predictive_value: 98.69% +/- 1.86% (micro average: 98.61%) (positive class: 1)		
ConfusionMatrix:			**ConfusionMatrix:**		
True	**0**	**1**	**True**	**0**	**1**
0	4166	301	**0**	4367	164
1	162	262	**1**	5	355

k-NN algoritmasında k parametresi, modelin doğruluğunu doğrudan etkileyen kritik bir hiperparametredir. Literatürde genellikle k değerinin 3 ile 10 arasında değiştiği görülmektedir. Bu çalışmada, farklı k değerleri (k=3, 5, 7, 9) denenmiş ve her biri için çapraz doğrulama ile sınıflandırma başarımı karşılaştırılmıştır. Elde edilen sonuçlara göre, $k = 3$ değeri en yüksek doğruluk oranını sağlamış ve bu nedenle modelde $k = 3$ olarak belirlenmiştir. Daha yüksek k değerlerinde modelin doğruluğunun azaldığı veya sınıflar arası ayrımın belirsizleştiği gözlemlenmiştir. Bu bağlamda, $k = 3$ seçimi hem literatürle uyumlu hem de deneysel olarak en iyi performansı veren değer olmuştur. Yapılan değerlendirme sonucunda, k-NN algoritmasının sınıf bazlı başarımı incelenmiştir. Karmaşıklık matrisi (Confusion matrix) ve sınıflandırma metrikleri üzerinden gerçekleştirilen analizde, özellikle "Sıcaklık Değişimi" kategorisinde modelin performansının diğer sınıflara kıyasla belirgin biçimde düşük olduğu gözlemlenmiştir. Bu sınıf için elde edilen precision (%61,84) ve recall (%46,54) değerleri, modelin hem doğru pozitifleri kaçırdığını hem de bu sınıfa ait olmayan örnekleri yanlış şekilde bu sınıf olarak etiketlediğini göstermektedir. Ayrıca dikkat çeken bir diğer bulgu, "Sıcaklık Değişimi" ve "Tank Kalibrasyonu" kategorilerinin birbirine semantik olarak benzerlik göstermesi nedeniyle analizlerde sıklıkla karıştırılıyor olmasıdır. Bu karışıklık, özellikle "Sızıntı" gibi kritik kategorilerin gözden kaçmasına ve analiz sürecinde önemli risklerin atlanmasına yol açmaktadır. Bu durum, modelin genelleme kapasitesini sınırlamakta ve saha uygulamaları açısından güvenilirliği azaltmaktadır. Söz konusu karışıklığın önüne geçebilmek için daha ayırt edici özelliklerin tanımlanması, gerekirse bu iki kategori için sınıf ayrıştırma amacıyla ikincil modellerin veya ön işleme tekniklerinin kullanılması önerilmektedir. Modelin sonuçları farklı başarı ölçütleri (TP, kesinlik, hassasiyet, F ölçütü, ROC değeri) hesaplanmış olup sonuçlar Tablo.5'de gösterilmiştir. k-NN algoritmasına göre arıza kategorisinin sınıflandırma başarısı yüksekken en düşük başarıya sahip kategorinin sıcaklık değişimi olduğu görülmüştür.

Tablo 5. k-NN Algoritması Performans Ölçütü Sonuçları

	Gerçek Pozitif (TP)	**Kesinlik (Precision)**	**Hassasiyet (Recall)**	***F*-Ölçütü**	**ROC Değeri**
Tank Kalibrasyonu	0,678	0,6779	0,6027	0,638	0,806
Dolum-Tesisat	0,802	0,8018	0,8995	0,847	0,867
Sıcaklık Değişimi	0,618	0,6184	0,4654	0,529	0,850
Arıza	0,963	0,9638	0,9989	0,981	0,878

Tablo 6. Rastgele Orman Algoritması Karmaşıklık Matrisi

Rastgele Orman Algoritması Kalibrasyon Modeli Karmaşıklık Matrisi			**Rastgele Orman Algoritması Dolum-Tesisat Modeli Karmaşıklık Matrisi**		
positive_predictive_value: 88.80% +/- 2.17% (micro average: 88.84%) (positive class: 1)			positive_predictive_value: 98.75% +/- 0.80% (micro average: 98.75%) (positive class: 1)		
ConfusionMatrix:			**ConfusionMatrix:**		
True	**0**	**1**	**True**	**0**	**1**
0	3486	159	**0**	2316	17
1	139	1107	**1**	32	2526
Rastgele Orman Algoritması Sıcaklık Değişimi Modeli Karmaşıklık Matrisi			**Rastgele Orman Algoritması Arıza Modeli Karmaşıklık Matrisi**		
positive_predictive_value: 81.09% +/- 3.35% (micro average: 80.99%) (positive class: 1)			positive_predictive_value: 97.99% +/- 0.61% (micro average: 97.99%) (positive class: 0)		
ConfusionMatrix:			**ConfusionMatrix:**		
True	**0**	**1**	**True**	**0**	**1**
0	4217	90	**0**	430	31
1	111	473	**1**	89	4341

Rastgele orman algoritmasında analize başlamadan önce rastgele orman algoritmasının özelliklerini tanımlamamız gerekmektedir. Random Forest algoritması için hiperparametre optimizasyonu gerçekleştirilmiştir. Bu kapsamda, modelin başarımını artırmak amacıyla oluşturulacak ağaç sayısı (n_estimators) 400 olarak, her ağacın maksimum derinliği (max_depth) ise 30 olarak belirlenmiştir. Bu değerler, modelin genelleme yeteneğini korurken daha yüksek doğruluk sağlaması amacıyla deneysel olarak seçilmiştir. Yapılan bu iyileştirme, modelin aşırı öğrenme riskinden uzak, daha dengeli sonuçlar üretmesine katkı sağlamıştır.Yapılan değerlendirme sonucunda, Rastgele Orman algoritmasının sınıf bazlı başarımı detaylı bir şekilde incelenmiştir. Confusion matrix ve sınıflandırma metrikleri üzerinden yapılan analizde, özellikle "Sıcaklık Değişimi" kategorisinin, diğer kategorilere kıyasla belirgin biçimde düşük performans sergilediği gözlemlenmiştir. Bu sınıf için elde edilen precision (%81,09) ve recall (%84,02) değerleri, modelin doğru pozitifleri yeterince yakalayamadığını ve aynı zamanda bazı yanlış pozitifler ürettiğini göstermektedir. Özellikle, "Sıcaklık Değişimi" kategorisinde yüksek false positive (yanlış pozitif) oranı, modelin pozitif sınıfları doğru sınıflandıramadığını ve bu kategoriyi gereksiz yere diğer kategorilerle karıştırdığını ortaya koymaktadır. Diğer dikkat çeken bir bulgu ise "Sıcaklık Değişimi" ve "Tank Kalibrasyonu" kategorilerinin birbirine benzerlik gösterdiği ve bu benzerlik nedeniyle modelin bu iki kategoriyi sıklıkla karıştırdığıdır.Tank Kalibrasyonu ve Dolum-Tesisat kategorileri ise oldukça yüksek başarı ile sınıflandırılmıştır. Tank Kalibrasyonu kategorisinde elde edilen precision (%88,80) ve recall (%87,40) değerleri, modelin doğru pozitifleri çoğunlukla doğru şekilde sınıflandırabildiğini göstermektedir. Benzer şekilde, Dolum-Tesisat kategorisinde elde edilen precision (%98,75) ve recall (%99,33) değerleri de modelin bu kategoriyi çok doğru bir şekilde tanıyıp etiketlediğini göstermektedir.Modelin performansı farklı başarı ölçütleri (TP, kesinlik, duyarlılık, F ölçütü, ROC değeri) kullanılarak değerlendirilmiş ve sonuçlar Tablo.7'da sunulmuştur. Rastgele Orman algoritmasına göre arıza kategorisinin sınıflandırma başarısı yüksekken en düşük başarıya sahip kategorinin sıcaklık değişimi olduğu görülmüştür. Random Forest modelinin tank kalibrasyon hatalarını belirleme stratejisi, Gradient Boosted modeline kıyasla veriyi daha dengeli ve "oran" odaklı işlemektedir: Sınıflandırmada 1 tam ağırlık ile Difference (Fark) değişkeni en belirleyici parametre olarak öne çıkmaktadır. Gradient Boosted modelinden farklı olarak Rate (%) (0,8698) özniteliği çok yüksek bir öneme sahiptir. Bu durum, Random Forest modelinin kalibrasyon hatalarını sadece miktar farkıyla değil,

zamana yayılan değişim oranıyla da analiz ettiğini gösterir. Open (0,7099), Reduction (0,6166) ve Close (0,5948) değişkenleri güçlü birer yardımcı sinyal olarak karar sürecine dahil olmaktadır. EL (0,1546) ve Refuel (0,1325) değişkenleri sınırlı katkı sunarken, Oil Type hiçbir etki göstermemiştir Random Forest modelinin sıcaklık değişimlerini sınıflandırma stratejisi, fark ve azalış miktarını merkeze almaktadır: Sınıflandırmada 1 ağırlık ile Difference (Fark) değişkeni en belirleyici parametredir. Reduction (0,7397) ve Open (0,5923) öznitelikleri, sıcaklık kaynaklı hacimsel hareketleri tanımlamada güçlü yardımcı sinyallerdir. Rate (%) (0,5392) özniteliği anlamlı bir ağırlığa sahiptir; bu da modelin zamana bağlı değişim oranını dikkate aldığını gösterir. ell (0,3807) ve Refuel (0,3208) değişkenlerinin katkısı sınırlı kalırken, Oil Type (0,0000) hiçbir etki göstermemiştir. Random Forest modelinin dolum tesisat kaynaklı sorunları sınıflandırma stratejisi, operasyonel değişkenleri daha geniş bir yelpazede değerlendirmektedir: Sınıflandırmada 1 ağırlık ile Reduction (Azalış) değişkeni en belirleyici parametredir Close (0,7895), Difference (0,7790) ve Open (0,7628) öznitelikleri birbirine yakın ve yüksek ağırlıklarla karar sürecine eşlik etmektedir. Sell (0,3766) ve Rate (%) (0,2821) değişkenleri Diğer modellerin aksine, bu kategoride Refuel (Dolum) değişkeni 0,0000 değeriyle karar sürecinde hiçbir rol oynamamıştır. orta düzeyde bir etkiye sahipken, Oil Type ve SEL değişkenlerinin katkısı oldukça düşüktür. Random Forest modelinin arıza kategorisindeki sınıflandırma hiyerarşisi, operasyonel süreçlerin başlangıç ve bitiş verilerine odaklanmaktadır: Sınıflandırmada Open (1,0000) ve Close (0,9552) öznitelikleri en yüksek ağırlığa sahip olup, arızaların vardiya veya işlem döngüsü verileriyle doğrudan ilişkili olduğunu göstermektedir. Difference (0,9331), Rate (%) (0,8173) ve Reduction (0,7698) değişkenleri birbirine yakın yüksek ağırlıklarla karar mekanizmasına güçlü katkı sunmaktadır. Sell (0,3305) ve Tank No (0,2894) sınırlı bir etkiye sahipken, Refuel değişkeni 0,0000 değeriyle bu kategoride etkisiz kalmıştır.

Tablo 7. Rastgele Orman Algoritması Performans Ölçütü Sonuçları

	Gerçek Pozitif (TP)	**Kesinlik (Precision)**	**Hassasiyet (Recall)**	***F*-Ölçütü**	**ROC Değeri**
Tank Kalibrasyonu	0,8880	0,8880	0,8740	0,8807	0,981
Dolum-Tesisat	0,9875	0,9875	0,9933	0,9904	0,998
Sıcaklık Değişimi	0,8109	0,8109	0,8402	0,8246	0,977
Arıza	0,9799	0,9799	0,9929	0,9864	0,983

Tablo 8. Gradient Boosted Trees Algoritması Karmaşıklık Matrisleri

Gradient Boosted Algoritması Kalibrasyon Modeli Karmaşıklık Matrisi			**Gradient Boosted Algoritması Dolum-Tesisat Modeli Karmaşıklık Matrisi**		
positive_predictive_value: 87.44% +/- 1.86% (micro average: 87.43%) (positive class: 1)			positive_predictive_value: 99.10% +/- 0.75% (micro average: 99.09%) (positive class: 1)		
ConfusionMatrix:			**ConfusionMatrix:**		
True	**0**	**1**	**True**	**0**	**1**
0	3467	167	**0**	2325	26
1	158	1099	**1**	23	2517

Gradient Boosted Algoritması Sıcaklık Değişimi Modeli Karmaşıklık Matrisi			**Gradient Boosted Algoritması Arıza Modeli Karmaşıklık Matrisi**		
positive_predictive_value: 82.56% +/- 3.97% (micro average: 82.30%) (positive class: 1) ConfusionMatrix:			positive_predictive_value: 90.70% +/- 4.75% (micro average: 90.48%) (positive class: 1)		
ConfusionMatrix:			**ConfusionMatrix:**		
True	**0**	**1**	**True**	**0**	**1**
0	4234	126	**0**	4326	82
1	94	437	**1**	46	437

Gradient Boosted Trees Algoritması, karar ağacı temeline dayanan bir algoritmadır. Algoritmanın en avantajlı durumlarından biri de her koşulda iyi sonuç vermesidir. Algoritmanın özelliklerinden biri de sonuç için oluşturulacak karar ağacı sayısıdır. Çalışmamızda analize başlamadan önce Gradient Boosting Trees algoritmasının temel özelliklerinin tanımlanması önemlidir. Bu çalışma kapsamında, modelin performansını en üst düzeye çıkarmak amacıyla hiperparametre optimizasyonu uygulanmıştır. Bu doğrultuda, oluşturulacak ağaç sayısı (n_estimators) 300, her bir ağacın maksimum derinliği (max_depth) ise 30 olarak belirlenmiştir. Bu hiperparametre değerleri, modelin hem öğrenme kapasitesini artırmak hem de aşırı öğrenme (overfitting) riskini minimize etmek amacıyla deneysel olarak seçilmiştir. Uygulanan bu ayarlamalar sayesinde model, daha dengeli ve yüksek doğruluk oranlarına sahip sonuçlar üretebilmiştir. Modelin performansı farklı başarı ölçütleri (TP, kesinlik, duyarlılık, F ölçütü, ROC değeri) kullanılarak değerlendirilmiş ve sonuçlar Tablo.8'de sunulmuştur. Yapılan değerlendirmeler sonucunda, Gradient Boosted Algoritması ile elde edilen modelin sınıf bazlı başarımı incelenmiştir. Confusion matrix ve sınıflandırma metrikleri üzerinden gerçekleştirilen analizde, özellikle Sıcaklık Değişimi kategorisinde modelin performansının diğer sınıflara kıyasla belirgin biçimde düşük olduğu gözlemlenmiştir. Bu sınıf için elde edilen precision (%82,56) ve recall (%77,63) değerleri, modelin hem doğru pozitifleri kaçırdığını hem de bu sınıfa ait olmayan örnekleri yanlış şekilde etiketlediğini göstermektedir. Ayrıca dikkat çeken bir diğer bulgu, Sıcaklık Değişimi ve Tank Kalibrasyonu kategorilerinin birbirine semantik olarak benzerlik göstermesi nedeniyle sıkça karıştırıldığıdır. Bu karışıklık, özellikle modelin her iki sınıfı ayırt etmede zorlandığını ve sonuçta doğru sınıflandırmaların azalmasına yol açtığını ortaya koymaktadır. Gradient Boosted Trees algoritmasına göre dolum-tesisat kategorisinin sınıflandırma başarısı yüksekken en düşük başarıya sahip kategorinin sıcaklık değişimi olduğu görülmüştür. Gradient Boosted Trees algoritmasının Tank Kalibrasyonu sınıflandırmasında kullandığı karar mekanizmasını anlamak amacıyla gerçekleştirilen öznitelik önem analizi, modelin ayırt edici gücünün hangi parametrelerde yoğunlaştığını sayısal olarak ortaya koymaktadır. Analiz sonuçlarına göre özniteliklerin normalize edilmiş ağırlıkları şu şekildedir: Model, en yüksek ayırt edici gücü Sell (1,0000), Difference (0,9365) ve Refuel (0,8094) özniteliklerine atamıştır. Bu durum, kalibrasyon hatalarının temel olarak satış ve dolum işlemleri sırasında oluşan stok farklarından tespit edildiğini kanıtlamaktadır. Reduction (0,3392) orta düzeyde bir etki gösterirken; Rate (%) (0,1789), Close (0,1132) ve Open (0,0136) gibi değişkenlerin katkısı sınırlı kalmıştır. Tank No (0,0037), Oil Type (0,00002) ve SEL (0,0000) değerleri, bu parametrelerin sınıflandırma üzerinde istatistiksel bir etkisinin olmadığını ortaya koymuştur. Dolum Tesisat kategorisi için öznitelik ağırlıkları incelendiğinde, modelin kararlarını neredeyse tamamen dolum sürecine dayandırdığı görülmektedir. Sınıflandırmada 1,0000 ağırlık ile Refuel (Dolum) değişkeni mutlak belirleyicidir. Difference (0,3635) değişkeni ikincil düzeyde etki ederken, Reduction (0,0392) ve Sell (0,0228) değişkenlerinin katkısı oldukça düşüktür. Rate (%), Oil Type, SEL ve Close gibi öznitelikler 0,00 veya buna yakın değerlerle karar sürecinde rol oynamamıştır. Tank Kalibrasyonu modelinin aksine, bu kategoride Refuel özniteliğinin tek başına domine edici olması, tesisat sorunlarının doğrudan dolum operasyonları sırasında karakterize

edildiğini kanıtlamaktadır. Sıcaklık Değişimi kategorisi için öznitelik ağırlıkları incelendiğinde, modelin kararlarını tanktaki hacimsel azalış hareketlerine dayandırdığı görülmektedir. Sınıflandırmada 1,0000 ağırlık ile Reduction (Azalış) değişkeni en belirleyici parametredir. Difference (0,5928) ve Open (0,2745) öznitelikleri ikincil düzeyde önem taşırken, Close (0,0843) ve Refuel (0,0483) değişkenlerinin etkisi oldukça sınırlıdır. Rate (%), SEL ve Oil Type gibi öznitelikler 0,00 değeriyle karar sürecinde hiçbir rol oynamamıştır. Arıza kategorisi için yapılan analizde, modelin operasyonel kapanış verilerini birincil sinyal olarak kullandığı görülmektedir: Sınıflandırmada 1 tam ağırlık ile Close (Kapanış) değişkeni mutlak belirleyicidir. Difference (0,4972) değişkeni ikincil düzeyde güçlü bir etkiye sahipken, Reduction (0,0780) ve Sell (0,0357) değişkenlerinin katkısı oldukça düşüktür. Rate (%) (0,0083), Tank No (0,0041) ve Open (0,0021) gibi öznitelikler çok düşük seviyede kalırken; Oil Type ve SEL değişkenleri 0,00 değeriyle karar sürecinde rol oynamamıştır. Arıza kategorisinde Close özniteliğinin domine edici olması, sistemdeki teknik arızaların genellikle vardiya veya işlem kapanış verilerindeki tutarsızlıklar üzerinden karakterize edildiğini göstermektedir.

Tablo 9. Gradient Boosted Trees Algoritması Performans Ölçütü Sonuçları

	Gerçek Pozitif (TP)	Kesinlik (Precision)	Hassasiyet (Recall)	F-Ölçütü	ROC Değeri
Tank Kalibrasyonu	0,8744	0,8744	0,8681	0,8707	0,980
Dolum-Tesisat	0,9910	0,9910	0,9898	0,9904	0,998
Sıcaklık Değişimi	0,8256	0,8256	0,7763	0,7988	0,966
Arıza	0,9070	0,9070	0,8420	0,8721	0,980

Tablo 10. ADABOOST Algoritması Karmaşıklık Matrisi

ADABOOST Algoritması Kalibrasyon Modeli Karmaşıklık Matrisi			**ADABOOST Algoritması Dolum-Tesisat Modeli Karmaşıklık Matrisi**		
positive_predictive_value: 82.60% +/- 3.44% (micro average: 82.51%) (positive class: 1)			positive_predictive_value: 96.80% +/- 0.70% (micro average: 96.80%) (positive class: 1)		
ConfusionMatrix:			**ConfusionMatrix:**		
True	**0**	**1**	**True**	**0**	**1**
0	3388	148	**0**	2265	35
1	237	1118	**1**	83	2508
ADABOOST Algoritması Sıcaklık Değişimi Modeli Karmaşıklık Matrisi			**ADABOOST Algoritması Arıza Modeli Karmaşıklık Matrisi**		
positive_predictive_value: 77.91% +/- 5.94% (micro average: 77.51%) (positive class: 1)			positive_predictive_value: 92.22% +/- 3.22% (micro average: 92.19%) (positive class: 1) ConfusionMatrix:		
ConfusionMatrix:			**ConfusionMatrix:**		
True	**0**	**1**	**True**	**0**	**1**
0	4196	108	**0**	4336	94
1	132	455	**1**	36	425

Modelin sonuçları farklı başarı ölçütleri (TP, kesinlik, hassasiyet, F ölçütü, ROC değeri) hesaplanmış olup sonuçlar Tablo.11'da gösterilmiştir. Model en iyi sonucu dolum-tesisat kategorisinde veriyorken en kötü sonucu sıcaklık değişimi kategorisinde verdiği görülmüştür.

Tablo 11. ADABOOST Algoritması Performans Ölçütü Sonuçları

	Gerçek Pozitif (TP)	Kesinlik (Precision)	Hassasiyet (Recall)	F-Ölçütü	ROC Değeri
Tank Kalibrasyonu	0,826	0,826	0,8831	0,853	0,972
Dolum-Tesisat	0,968	0,968	0,9862	0,977	0,996
Sıcaklık Değişimi	0,7791	0,7791	0,808	0,7921	0,967
Arıza	0,9222	0,9222	0,8187	0,8668	0,965

Tablo 12. Karar Ağacı (J48) Algoritması Karmaşıklık Matrisler

Karar Ağacı (J48) Algoritması Kalibrasyon Modeli Karmaşıklık Matrisi			**Karar Ağacı (J48) Algoritması Dolum-Tesisat Modeli Karmaşıklık Matrisi**		
positive_predictive_value: 87.13% +/- 2.25% (micro average: 87.07%) (positive class: 1)			positive_predictive_value: 98.44% +/- 0.91% (micro average: 98.43%) (positive class: 1)		
ConfusionMatrix:			**ConfusionMatrix:**		
True	**0**	**1**	**True**	**0**	**1**
0	3460	155	**0**	2308	36
1	165	1111	**1**	40	2507
Karar Ağacı (J48) Algoritması Sıcaklık Değişimi Modeli Karmaşıklık Matrisi			**Karar Ağacı (J48) Algoritması Arıza Modeli Karmaşıklık Matrisi**		
positive_predictive_value: 80.88% +/- 5.99% (micro average: 80.60%) (positive class: 1)			positive_predictive_value: 93.32% +/- 5.04% (micro average: 93.07%) (positive class: 1)		
ConfusionMatrix:			**ConfusionMatrix:**		
True	**0**	**1**	**True**	**0**	**1**
0	4218	106	**0**	4340	89
1	110	457	**1**	32	430

Algoritmanın özelliklerinden biri sınıflandırılacak olan verinin hangi özelliğinin kullanarak dallandırılacağının belirlenmesidir. Bunun için özellik seçim ölçütlerinden bilgi kazanımı (Information Gain) yararlanır, ancak herhangi bir özelliğin aldığı değerin fazla çeşitli olması, kriteri, bu özelliğin kullanılarak dallanmaya zorlaması bir taraflılık oluşturduğundan GINI indeksinin kullanılması daha yaygın bir seçimdir. Çalışmamızda J48 algoritmasında özellik seçimi ölçütlerinden GINI indeksinden yararlanılmıştır. Modelin sonuçları farklı başarı ölçütleri (TP, kesinlik, hassasiyet, F ölçütü, ROC değeri) hesaplanmış olup sonuçlar Tablo.13'da gösterilmiştir. Model en iyi sonucu dolum-tesisat kategorisinde veriyorken en kötü sonucu sıcaklık değişimi kategorisinde verdiği görülmüştür.

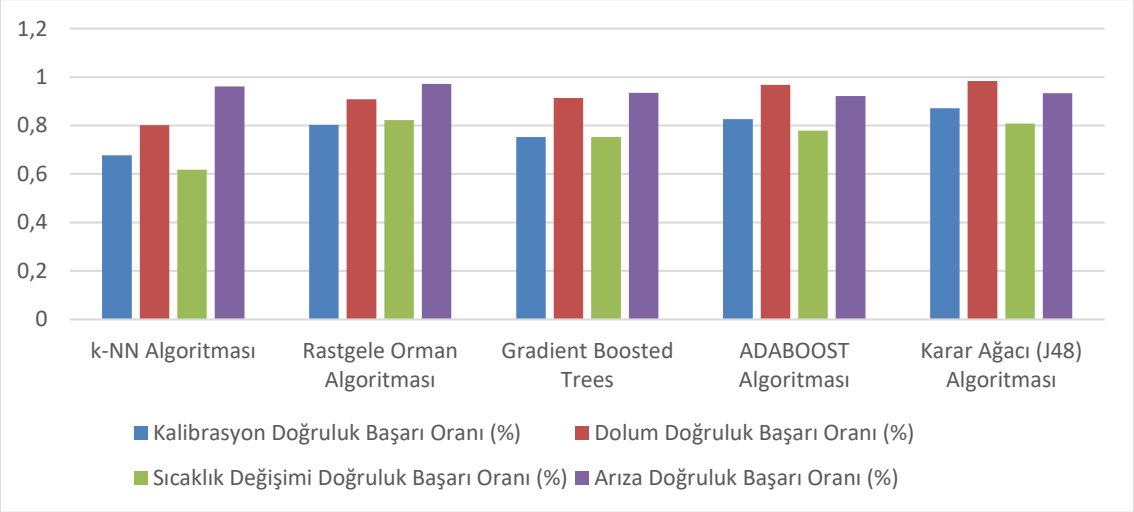
Tablo. 13 Karar Ağacı (J48) Algoritması Başarı Sonuçları

	Gerçek Pozitif (TP)	Kesinlik (Precision)	Hassasiyet (Recall)	F-Ölçütü	ROC Değeri
Tank Kalibrasyonu	0,8713	0,8713	0,8776	0,874	0,925
Dolum-Tesisat	0,9844	0,9844	0,9858	0,9851	0,989
Sıcaklık Değişimi	0,8088	0,8088	0,8117	0,809	0,913
Arıza	0,9332	0,9332	0,8285	0,8768	0,916

Tablo.14 Algoritmaların Başarı Oranları

	Kalibrasyon Doğruluk Başarı Oranı (%)	Dolum Doğruluk Başarı Oranı (%)	Sıcaklık Değişimi Doğruluk Başarı Oranı (%)	Arıza Doğruluk Başarı Oranı (%)
***k*-NN Algoritması**	0,677	0,801	0,618	0,961
Rastgele Orman Algoritması	0,803	0,908	**0,822**	**0,972**
Gradient Boosted Trees Algoritması	0,753	0,914	0,752	0,935
ADABOOST Algoritması	0,826	0,968	0,779	0,922
Karar Ağacı (J48) Algoritması	**0,871**	**0,984**	0,808	0,933

Tablo.15 Algoritmaların Başarı Oranları (Sütun Grafiği)



6. TARTIŞMA VE SONUÇ

Türkiye içerisinde faaliyet gösteren bir petrol şirketinin belirli bölgelerinde hizmet veren istasyonlarından 2022-2023 tarihleri arasındaki 4 ana kategori için gerçek veriler kullanılarak çeşitli veri madenciliği yöntemleri eğitilerek test edilmiştir. Test verilerinin kesinlik, hassasiyet, F ölçütü ve ROC değerleri performans ölçütleri ile ölçülmüş olup algoritmaların başarısı değerlendirilmiştir. Tablo.14 ve Tablo.15'de 4 ayrı kategorinin çeşitli algoritmalar ile başarı oranları gösterilmiştir. Söz konusu tablo'da özetlenen sonuçlara göre kalibrasyon, dolum, sıcaklık değişimi ve arıza kategorilerinin k-NN algoritması dışında diğer yöntemlerde iyi sonuç verdiği görülmüştür. Sıcaklık Değişimi sınıfı, açık ara en zorlayıcı kategori olarak öne çıkmakta ve model performansını ciddi şekilde düşürmektedir. Söz konusu performans düşüklüğünün, veri seti dağılımı dengeli olmasına rağmen, 'Sıcaklık Değişimi' ve 'Tank Kalibrasyonu' sınıflarının benzer öznitelik profillerine (feature similarity) sahip olmasından kaynaklandığı değerlendirilmektedir. Yapılan analizler, bu iki kategorinin ayırt edilmesinde kullanılan parametrelerin birbirine yakın değerler üretmesi nedeniyle algoritmaların karar sınırlarını (decision boundaries) belirlemekte zorlandığını göstermiştir. Özellikle Tank Kalibrasyonu ile olan sınıf karışıklığı dikkate alınarak, bu iki sınıf için ikincil sınıflandırıcıların ya da daha gelişmiş

ayrıştırma özelliklerinin (feature selection) geliştirilmesi önerilmektedir. AdaBoost ve Rastgele Orman algoritmaları, genel performans açısından diğer modellere kıyasla daha başarılı sonuçlar vermiştir. k-NN algoritması, genelleme kabiliyeti zayıf olması nedeniyle tüm sınıflarda görece düşük performans göstermiştir. İşletmeye son kontrollerin yapılması adına model önerisi yapılmış ve kullanılması önerilmiştir Modellerin karar mekanizmalarını daha iyi açıklayabilmek adına gerçekleştirilen özellik önemi (feature importance) analizi sonucunda, özellikle değişim hızı ve fark temelli özniteliklerin sınıflandırmada kritik rol oynadığı belirlenmiştir. Bu bağlamda, gelecekteki çalışmalarda bu iki sınıfı ayrıştıracak daha spesifik öznitelik mühendisliği (feature engineering) çalışmalarına odaklanılması hedeflenmektedir. Gelecek çalışmalarda, geliştirilecek yazılımlar sayesinde ve yapay zeka teknolojilerinin de bu alanda kullanılmaya başlanmasıyla veri tabanlarına yansıyan verilerin sorunsuz ve hızlı bir şekilde analizlerinin gerçekleştirilerek daha hızlı ve doğru aksiyonların alınması sağlanılabilir. Bu sayede, maliyetlerin düşürülmesi, üst yönetimin ve devlet kurumlarının doğru bilgilendirilmesi sağlanacak, ileriki dönemlerde alınması gereken tedbirlerin daha etkili olması sağlanacaktır. Geliştirilen modeller, istasyon bazlı anomali tespit sistemlerinin oluşturulmasına altyapı sağlayabilir. Böylece potansiyel kalibrasyon hataları, dolum kaynaklı sapmalar veya sıcaklık etkileri erken aşamada belirlenerek müdahale süresi kısaltılabilir. Bu durum hem operasyonel kayıpları azaltacak hem de bakım planlamasının daha etkin yapılmasına katkı sağlayacaktır. Çalışma, yapay zekâ destekli karar destek sistemlerinin akaryakıt sektörüne entegrasyonunun mümkün ve uygulanabilir olduğunu göstermektedir. Gelecekte geliştirilecek yazılımlar sayesinde veri tabanına yansıyan ölçüm verileri anlık olarak analiz edilebilir, otomatik uyarı sistemleri oluşturulabilir ve müdahale süreçleri hızlandırılabilir. Bu durum sektörde dijital dönüşümün hızlanmasına katkı sağlayacaktır. Sonuç olarak, bu çalışma yalnızca algoritmik performans karşılaştırması sunmamakta; aynı zamanda akaryakıt sektöründe veri temelli operasyon yönetiminin uygulanabilirliğini ortaya koymaktadır. Özellikle Sıcaklık Değişimi ve Tank Kalibrasyonu sınıfları arasındaki ayrımın güçlendirilmesi için ileri düzey öznitelik mühendisliği, hibrit modelleme yaklaşımları ve derin öğrenme temelli yöntemler gelecekteki çalışmalar için önemli araştırma alanları olarak değerlendirilmektedir. Ayrıca gerçek zamanlı veri akışı ile entegre çalışan, otomatik öğrenen ve kendini güncelleyebilen sistemlerin geliştirilmesi, sektörde sürdürülebilir performans artışı sağlayabilecek önemli bir adım olacaktır.

ÇIKAR ÇATIŞMASI

Yazarlar, bilinen herhangi bir çıkar çatışması veya herhangi bir kurum/kuruluş ya da kişi ile ortak çıkar bulunmadığını onaylamaktadırlar

YAZAR KATKISI

Bu çalışmada Sabahattin Mert Berkmen ve Sabahattin Kerem Aytulun, çalışmanın kavramsal ve tasarım süreçlerinin belirlenmesi çalışmanın kavramsal ve tasarım süreçlerinin yönetimi, veri analizi ve yorumlama, makale taslağının oluşturulması, fikirsel içeriğin eleştirel incelemesi; Sabahattin Mert Berkmen kavramsal ve tasarım süreçlerinin belirlenmesi, veri toplama, veri analizi ve yorumlama, makale taslağının oluşturulması, fikirsel içeriğin eleştirel incelemesi başlıklarında katkı sunmuşlardır

KAYNAKLAR

1. Agrawal, R., Imielinski, T. ve Swami, A. (1993) Database mining: A performance perspective. *IEEE Transactions on Knowledge and Data Engineering*. 5(6), 914-925. doi: 10.1109/69.250074
2. Aydın C. (2019) Makine öğrenmesi algoritmaları kullanılarak itfaiye istasyonu ihtiyacının sınıflandırılması. Avrupa Bilim ve Teknoloji Dergisi, 14(1), 169-175 doi: https://doi.org/10.31590/ejosat.458613
3. Boland, J., Baumann, D. ve Dziegielewski, B. (1981) *An assessment of municipal and industrial water use forecastingapproaches*. Defence Technıcal Informatıon Center, Virginia. Erişim adresi: https://apps.dtic.mil
4. Chen, M., Huang, C. ve Wu, P. (2005) Aggregation of orders in distribution centers using data mining. *Expert Systems with Applications*, 28(3), .453-460 doi:https://doi.org/10.1016/j.eswa.2004.12.006
5. Coşkun, C., Baykal, A. (2011). Veri madenciliğinde sınıflandırma algoritmalarının bir örnek üzerinde karşılaştırılması . *XII. Akademik Bilişim Konferansı*, 2-4 Şubat, Malatya. Erişim adresi: https://ab.org.tr/ab11/kitap/coskun_baykal_AB11.pdf
6. Dalman, A, A.(2017). Wet-stock management and leak detecıtıon system for fuel tanks. Yüksek Lisans Tezi, Yeditepe Üniversitesi. Erişim adresi: https://tez.yok.gov.tr
7. Dominic, D. ve Dagbui, A. (2014) Dealing with construction cost overruns using data mining. *Construction Management and Economics*, 32(7-8). 682-694 doi: 10.1080/01446193.2014.933854
8. Demirel, Ş. ve Yakut, G. (2019).Karar ağacı algoritmaları ve çocuk işçiliği üzerine bir uygulama. *Sosyal Bilimler Araştırma Dergisi*, 8(4). 52-65. Erişim adresi: https://www.acarindex.com/sosyal-bilimler-arastirma-dergisi/karar-agaci-algoritmalari-ve-cocuk-isciligi-uzerine-bir-uygulama-1116043
9. Dudas, C., Ng, A., Pehrsson, L. ve Boström, H. (2013). Integration of data mining and multi-objective optimisation for decision support in production systems development. *International Journal of Computer Integrated Manufacturing*, vol.27(9). 824-839 doi: https://doi.org/10.1080/0951192X.2013.834481
10. Doğan, E. K. ve Şentürk, A.(2021). Veri madenciliği yöntemleri ile işveren sektörünün sınıflandırılması. *Avrupa Bilim ve Teknoloji Dergisi*, (32). 227-234 doi: https://doi.org/10.31590/ejosat.1039844
11. Dos santos, B. S., Steiner, M. T., Fenerich, A. T. ve Lima, R. H.(2019). Data mining and machine learning techniques applied to public health. *Computers & Industrial Engineering,* 13(2), .doi: https://doi.org/10.1016/j.cie.2019.106120
12. Dönmez, Z. S. (2008).Bayi performans değerlendirilmesinde bir veri madenciliği uygulaması. Yüksek Lisans Tezi, İstanbul Teknik Üniversitesi. Erişim adresi: https://tez.yok.gov.tr
13. Kuşaksızoğlu,B. (2006). Veri madenciliği yardımıyla mobil telekomünikasyon şebekelerinde sahtekarlık tespiti. Yüksek Lisans Tezi, Bahçeşehir Üniversitesi. Erişim adresi: https://tez.yok.gov.tr
14. Özekes, S. (2003).Veri madenciliği modelleri ve uygulama alanları, *İstanbul Ticaret Üniversitesi Dergisi*, 2(3). 65-82. URL= https://izlik.org/JA97FN46RZ
15. Schuh, G., Prote, J. P. ve Hunnekes, P. (2020).Data mining methods for macro level process planning. *Procedia CIRP*,(88), 48-53 doi: https://doi.org/10.1016/j.procir.2020.05.009
16. Sforna, M.(2000). Data mining in a power company customer database. *Electric Power Systems Research*, 55(3). 201-209 doi: https://doi.org/10.1016/S0378-7796(00)00086-9

17. Şimşek, F. (2019).Veri madenciliği sınıflandırma tekniklerini kullanarak üretim sistemlerinde hatalı ürünlerin tespit edilmesi. Yüksek lisans tezi, Kırıkkale Üniversitesi. Erişim adresi: https://tez.yok.gov.tr

18. Vezhnevets, A. ve Vezhnevets, V.(2005). 'Modest adaboost' – teaching adaboost to generalize better. *Graphicon*, 12(5). 987-997 url=https://api.semanticscholar.org/CorpusID:16793346