

Araştırma Makalesi

Keşifçi Veri Analizi ile Eczane Satış Analizi ve Satış Tahmini

Pharmacy Sales Analysis and Sales Forecasting with Exploratory Data Analysis

Mansur BEŞTAŞ

Dr. Siirt, Türkiye

mansur@siirt.edu.tr

<https://orcid.org/0000-0002-8192-2044>

Makale Geliş Tarihi	Makale Kabul Tarihi
29.11.2022	10.03.2023

Öz

İnsan sağlığı açısından önemli bir yeri olan ilaçların günlük kullanımının ve erişilebilirliğinin artışı ile tüketimi artmaya başlamıştır. Tüm işletme türlerinde olduğu gibi ilaç sektörünün son kullanıcı ile temasının olduğu eczanelerde satışların bilinmesi ve geleceğe yönelik öngörü geliştirilmesi önemlidir. İlaç sektörünün bireylere yönelik gerçekleşen satışlarının önemli bir kısmı ilaçlardan oluşmaktadır. İlaç satışları, tüketim ürünlerinden farklı bir konumda bulunmaktadır. Tüketim ürünleri, ihtiyacının yanı sıra gelişen sosyo-ekonomik durum ve ürün tutundurma faaliyetlerinden etkilenmekte iken sağlık ürünleri ağırlıklı olarak zorunlu hallerde gerçekleşen bir tüketim ürünüdür. Sağlık sektörü alanında tüketilen ürünler keşifsel analiz ile incelendiğinde ilgili yerelde halk sağlığına yönelik değişimin ortaya çıkarılmasını sağlar. İşletmelerin verimliliği açısından gelecekteki satışların mümkün mertebede tahmin edilmesine bağlıdır. Bu çalışmada Güneydoğu Anadolu bölgesinde bulunan bir eczanenin yaklaşık üç yıllık satış verileri keşifçi veri analizi ile incelenmiş ardından makine öğrenmesi ve zaman serileri algoritmaları ile sonraki 15 güne ait satış verilerinin tahmin edilmeye çalışılmıştır. Elde edilen en başarılı sonuç ARIMA yöntemi ile ortalama hataların karelerinin karekök değeri (RMSE) 23 olarak elde edilmiştir.

Anahtar Kelimeler: Veri Madenciliği, İlaç Satış Tahmini, Makine öğrenmesi, Zaman Serileri, Veri Analizi

Abstract

With the increase in daily use and accessibility of drugs, which have an important place in human health, their consumption has started to increase. As in all business types, it is important to know sales and to develop foresight for the future in pharmacies where the pharmaceutical industry is in contact with the end user. Drugs constitute a significant part of the sales of the pharmaceutical industry to individuals. Pharmaceutical sales are in a different position from consumer products. While consumer products are affected by the developing socio-economic situation and product promotion activities as much as their needs, health products are mainly consumption products that are made in obligatory situations. When the products consumed in the field the health sector are examined with exploratory analysis, it provides to reveal the change in public health in the relevant region. In terms of the efficiency of businesses, it depends on predicting future sales as much as possible. In this study, approximately three years of sales data of a pharmacy located in the Southeastern Anatolia region were examined with exploratory data analysis, and then the sales data for the next 15 days were tried to be estimated by machine learning and time series algorithms. The most successful result was obtained with the ARIMA method with a root mean square error (RMSE) of 23.

Keywords: Data Mining, Drug Sales Prediction, Machine learning, Time Series, Data Analysis

1. Giriş

Teknolojinin gelişimi ile birlikte verinin kayıt altına alınması kolaylaştırmıştır. Teknolojinin etkilediği birçok alan bulunmaktadır. Bu alanlardan biri de sağlık alanıdır. Sağlık sektöründe verinin kayıt altına

Önerilen Atıf /Suggested Citation

Beştaş, M., 2023 Keşifçi Veri Analizi ile Eczane Satış Analizi ve Satış Tahmini, *Üçüncü Sektör Sosyal Ekonomi Dergisi*, 58(1), 765-782

alınması gittikçe iyileşen bir eğriye sahip olmasına karşın kayıt altına alınan verilerin barındırdığı faydalı bilgilerin ortaya çıkarılması ve kullanılması görece yetersizdir. Veriler ancak daha önceden belirlenmiş bir hedef doğrultusunda kullanımı ile değerli hale gelebilmektedir (Aydemir ve Yavuz, 2019; Kalikov, 2006).

Verilerin kullanılacağı alanlardan biri satış ve pazarlama alanıdır. Satın alma süreçlerinin kolaylaştığı günümüzde satış rakamlarını tahmin etmek ve bu minvalde stok tutmak gerek büyük gerekse küçük işletmelerde büyük öneme sahiptir. İşletmelerin müşterilerine ve paydaşlarına gerekli olan kaliteli bir satış deneyimi sunabilmesi, mümkün mertebede geleceği tahmin etmesine dayanmaktadır. Kısa veya uzun vadeli süreçler göz önünde bulundurularak satış miktarının kestirilememesi işletmenin performans açısından başarısız olmasına neden olmaktadır. Gelecekte beklenen satışların tahmin başarı oranı sadece işletmenin değil, işletme ile bağlantılı tedarik zincirinin başarısını da etkilemektedir. Satış miktarlarının tahmin edilmesi önemli olduğu kadar kar marjlarının artırılmasında en ucuz yöntemlerinden biridir. İşletmelerin gerek iç gerekse dış risklerden etkileme oranını azaltmaktadır.

Medikal ürünlerin satışı analizi ve satış tahmini, ilaç endüstrisinin stok yönetimi, pazarlama stratejileri ve genel iş planlaması hakkında bilinçli karar alması için önemli bir görevdir. Keşifsel veri analizi (KVA), son yıllarda büyük veri kümelerindeki kalıpları, ilişkileri ve anormallikleri belirlemek için yaygın olarak kullanılmaktadır.

KVA teknikleri kullanılarak yapılan medikal satış analizleri birçok araştırmacının ilgi konusu olmuştur. Uzun yıllar temel ilaç kategorisinde görülen CO-Trimoksazol, Amoksisilin, Klotrimazol, Aminofilin ve Rifampin ilaçlarının aylık tüketimi tahmin edilmesi için yapılan çalışmada en uygun sonuçların Yapay sinir ağı v Karar ağacı algoritmaları aracılığı ile elde edildiği görülmüştür (Ghousi, Mehrani, Momeni ve Anjomshooa, 2012). Bir başka çalışmada Okacet ve Stamlo Beta ilaçlarının tüketimi değerlendirmeye alınmıştır. Okacet, mevsimsel etkiye bağlı olarak dalgalanmadan etkilendirken Stamlo Beta mevsimsel etkiye maruz kalmamaktadır. Elde edilen sonuçlar, Winter üstsel düzleştirme modeli mevsimlik ilaçlarda verimli çalışır iken hareketli ortalamanın mevsimsel olmayan ilaçlarda daha doğru sonuçlar ürettiği görülmüştür (Anusha, Alok ve Shaik, 2014). Van Boeckel ve ark. (2014), antibiyotik ilaç tüketimini incelemiş ve ülkelerarasında antibiyotik tüketiminin, buna bağlı antibiyotik direncini ve geleceğe yönelik tahmini tüketim projeksiyonu ortaya koymuştur. Papanagnou ve Matthews-Amune (2018), çevrimiçi kaynaklar kullanılarak elde edilen büyük veri aracılığı ile perakende satışı yapan eczanelerde oluşan talep dalgalanmalarını ele almıştır.

Başka bir çalışma, IMS RPM® (Bölgesel İlaç Pazarı) Haftalık Veri tabanından alınan verileri kullanarak 2019 ve 2020'de toptancılardan psikiyatrik ilaçların haftalık eczane alımlarını incelemiştir. Bu çalışmada istatistiksel bir sonuç elde edilememiştir. Bu çalışmanın en önemli bulgusu, COVID-19 tecrit resmi duyurusu ile haftalık psikiyatrik ilaç alımlarının artışı arasında tanımlayıcı bulgu olmamasıdır (Jacob, Bohlken ve Kostev, 2021). Pakistan'daki altı ilaçtan altı farklı ilaç sınıfının gerçek satış verileri üzerine gerçekleştirilen nicel araştırmada Otoresgresif Entegre Hareketli Ortalama (ARIMA) ve Holt's-Winter modelini birleştirilerek yapılan yeni hibrit model ile başarılı sonuçlar elde edilmiştir (Siddiqui, Azmat, Ahmed ve Kummer, 2022).

Bu çalışmada, bir işletmenin ilaç satış verileri keşifçi veri analiz ile incelenmiş ve ardından veri setine uygulanan ön işleme süreçlerinin kavramsal çerçevesi verilmiştir. Çalışmanın yöntem bölümünde veri setinin oluşturulan öz nitelikleri ve değerlendirmeye esas kriterler belirlenmiştir. Çalışmanın bulgular bölümünde keşifçi veri analizi ile elde edilen veri seti genel perspektifi ortaya koyulmuş, makine öğrenmesi algoritmaları ile elde edilen sonuçlar tablo halinde verilmiş ve seçili zaman serileri algoritmaları ile elde edilen bulgular gerek tablo gerekse grafikler ile ortaya koyulmuştur. Çalışmanın son bölümünde elde edilen bulguların değerlendirmesi yapılmış olup gelecekte yapılacak çalışmalara yönelik temennilerde bulunulmuştur.

2.Keşifçi Veri Analizi (KVA)

KVA, istatistiksel yaklaşım ve grafik gösterimi kullanarak verilerden örüntülerin tespiti ve değişkenler arasındaki ilişkileri anlama amaçlı içgörü elde etme sürecidir. Bu süreç verilerden anlamlı bilgilerin ortaya çıkarılması amacıyla tanımsal ve çıkarımsal istatistik başta olmak üzere görselleştirme tekniklerinin kullanıldığı bir dizi istatistiksel tekniği barındırır. KVA, özünde, araştırma sorularının

cevaplanması veya hipotezlerin test edilmesi amacıyla verilerin bir araya getirilmesi, düzenlenmesi ve analiz etmeyi kapsayan bilimsel bir yöntemdir. İlk olarak Tukey tarafından ortaya atılmış olsa da 1983 yılında yayınladığı bir başka çalışma ile bu alandaki görüşlerini detaylandırmıştır. Tufféry, Tukey tarafında ortaya koyulan esasları kullanarak veri madenciliği ve istatistiği birleştirmiştir (Tukey, 1977; Mosteller ve Tukey, 1977; Hoaglin, Mosteller ve Tukey, 1983; Tufféry, 2011).

KVA'nın ortaya çıkışından bu yana, araştırmacılar tarafından çeşitli görselleştirme teknikleri geliştirilmiştir. Günümüzde KVA, iş süreçleri için kritik bir öneme sahiptir. İş süreçlerinde doğrulanmaya muhtaç hipotezi belirleyen değişkenler, özellikler arasındaki gizli kalıpların ve ilişkilerin tanımlanmasını sağlar. KVA, sonucunda veriden elde edilen iç görüler ile stratejik iş kararlarının alınmasına yardımcı olmaktadır.

KVA'nın olmazsa olmazı görselleştirme araçlarıdır. Görselleştirme araçları, genom dizisi veri analizinde, meteorolojik verilerde, tahmine dayalı analizde olmak üzere bir çok çeşitli alanda uygulamaya sahiptir (Khan ve Khan, 2011; Pabinger ve ark., 2014; Rautenhaus ve ark., 2017; Liu, Wang, Liu ve Zhu, 2017). KVA'nın birçok alanda kullanılıyor olması kabul edilme oranını göstermektedir. Bu nedenle, işletmenin büyüklüğüne bakılmaksızın başarısı bir yönden elindeki verilerin analiz sonucuna bağlı olduğu anlamına gelmektedir.

Birçok araştırma alanı gibi, KVA'nın gerekliliği zaman içerisinde kendini göstermiştir. Veri hacminin zaman içerisinde çeşitlenmesi ve hızının artması bu alandaki araştırmacıların süreçleri optimize etmesine neden olmuştur. Bu sayede KVA ölçeklenebilir, esnek ve sağlam bir yöntem haline gelmiştir. Böylelikle medikal, telekom, imalat, gıda, enerji vb. birçok alanda işletmelere yardımcı olmaktadır (Godfrey, Gryz ve Lasek, 2016; Dey, Ashour, Shi, Fong, ve Simon Sherratt, 2017; Das ve Tripathi, 2018; Bro ve ark., 2002).

Gerçek hayattan elde edilen veriler genellikle içerisinde hatalar barındırmaktadır. Bu hatalar kimi zaman veri kaynağının hatalı çalışmasında olmakla beraber verinin saklama koşullarında kaynaklı olmaktadır. Veri içerisinde hatalı veya eksik bilginin bulunmasının yanı sıra veri model odaklı olmayabilmektedir. Bu nedenle verinin ön işlemeden geçirilmesi gerekmektedir.

2.1. Veri Ön İşleme

Veri temelinde oluşturulacak bir modelin temel amacı uygulanabilir ve geliştirilebilecek bir yapının gerçekleştirilmesidir. Verinin düzenlenmesi modelin oluşturulmasında kritik aşamalardan biridir (Çetin ve Yıldız, 2022). Veri setleri genellikle direkt kullanılabilecek yapıda değildir. Bu nedenle verinin temizlenmesi, dönüştürülmesi ve zenginleştirilmesi gerekmektedir. Veri ön işleme, ham verilerin temizlenmesini ve kolayca analiz edilebilecek bir biçime dönüştürülmesini içeren veri analizinin kritik bir adımdır. Bu adım önemlidir çünkü farklı kaynaklardan toplanan veriler, analiz sonuçlarının doğruluğunu ve geçerliliğini etkileyebilecek hatalar, eksik değerler, aykırı değerler veya tutarsızlıklar içerebilir. Veri ön işlemedeki ilk adım, verilerdeki hataları ve tutarsızlıkları belirleyip düzeltmeyi içeren veri temizlemedir. Bu süreç, kopyaların kaldırılmasını, eksik değerlerin doldurulmasını ve veri girişi hatalarının düzeltilmesini içerir. Bir sonraki adım, verilerin analiz için uygun bir biçime dönüştürülmesini içeren veri dönüştürmedir. Bu adım, verilerin ölçeklendirilmesini ve standartlaştırılmasını, veri dağılımının normalleştirilmesini ve kategorik değişkenlerin kodlanmasını içerebilir (Meeyai, 2016).

Verilerin ölçeklenmesi ve standardize edilmesi, tüm özelliklerin benzer bir ölçekte olması için verilerin dönüştürülmesini içerir. Bu önemlidir, çünkü bazı özellikler diğerlerinden daha geniş aralıklara sahip olabilir ve bu da bazı makine öğrenimi algoritmalarının performansını etkileyebilir. Veri dağılımının normalleştirilmesi, verilerin normal bir dağılım izleyecek şekilde dönüştürülmesini içerir. Bu önemlidir, çünkü birçok istatistiksel yöntem verilerin normal bir dağılım izlediğini varsayar (Soley-Bori, 2013).

Kategorik değişkenlerin kodlanması, cinsiyet veya meslek gibi kategorik verilerin analizde kullanılabilecek sayısal değerlere dönüştürülmesini içerir. Kategorik değişkenlerin kodlanması, makina öğrenmesi algoritmalarına uygun hale getirilmesidir. Algoritmaların daha verimli olarak çalışabilmesi amacıyla kimi zaman değerlerin 0-1 arasında ölçeklendirilmesi, 0-1 olarak iki kategoriye indirgenmesi, uç ve belirsiz değerlerin temizlenmesi ve benzeri yöntemlerin uygulanması gerekmektedir (Zhang, 2016).

Veriler temizlendikten ve dönüştürüldükten sonra, verilerdeki kalıpları ve ilişkileri ortaya çıkarmak için tanımlayıcı istatistikler, görselleştirmeler ve çıkarımsal istatistikler gibi veri içerisinde gizli kalmış bilgilerin ortaya çıkarılması için kullanılabilir.

KVA uyumlu veri ön işleme adımlarının sağlıklı olarak gerçekleştirilmesi, oluşturulacak makine öğrenmesi modelinin başarısında önemli etkiye sahiptir. Tablo 1’de veri ön işleme süreçleri verilmiştir.

Tablo 1: KVA Uyumlu Veri Ön İşleme Matrisi

Veri Keşfi			
Veri Kümesinin Şekli	Veri Özelliklerinin Türü	Eksik ve Benzersiz Veri Değerleri	
Merkezi Eğilim	Dağılım Ölçüsü	Birliktelik ve İlişki Ölçüsü	
Öz Nitelik Mühendisliği			
Standardizasyon	Normalizasyon	Kodlama	Kategorileştirme
Tekilleştirme	Vektörleştirme	Kişiselleştirme	
Binarization (sadece 1-0 olarak değerleri değiştirme, iki kategoriye bölme)	Encoding (belli etiketlere göre dönüşüm)	Imputation (uç ve belirsiz verilerin temizlenmesi)	
Öz Nitelik Seçimi			
Trend’e göre seçim	İlişkiye Göre Seçim	İlgiye Göre Seçim	Puanlamaya Göre Seçim
Bağımlı Değişken			
Sürekli Veri		Kategorik Veri	
Veri İndirgeme			
Gözlem Sayısı İndirgeme		Değişken Sayısı İndirgeme	

Belli başlı veri ön işleme teknikleri şunlardır:

Veri temizleme: Bu, verilerdeki hataları ve tutarsızlıkları tanımlamayı ve düzeltmeyi içerir. Veri seti içerisinde yinelenen veriler bulunuyor ise analiz sonuçlarında yanlışlığın ortaya çıkmaması adına bu veriler kaldırılabilir. Veri seti içerisinde bir veya birden fazla alan eksik değer barındırabilmektedir. Eksik veri değerlerinin olması durumunda ilgili veriler ortalama veya medyan atama vb. yöntemler ile doldurulabilir (Agarwal, 2013). Veri girişlerinden kaynaklı veri seti içerisinde hatalı durumlar gözlemlenebilir. Bu hatalar veri girişini yapan personelin hatasından kimi zaman geriyi kayıt altına alan internet nesnesi, sensör vb. cihazlardan kaynaklanabilmektedir. Bu hatalar gözden geçirilmeli ve hatanın

yapısına göre uygun değişiklikler ile veri düzeltilmelidir. Veri setinin barındırdığı ölçüm değerlerinin normal sınırları dışında kalan veya ölçüm sürecinde hayatın normal akışının dışında bir durumdan kaynaklı istisnai değerler olabilir. Bu değerler aykırı değer olarak adlandırılmaktadır. Aykırı değerler analiz sonuçlarını çarpıtabilmektedir. Doğru sonuçların sağlanması amacıyla aykırı veriler görmezden gelinebilir (Mughal, 2018).

Veri dönüştürme: Bu, verilerin analiz için uygun bir biçime dönüştürülmesini içerir. Verileri ölçeklendirmeyi ve standartlaştırmayı, veri dağılımını normalleştirmeyi ve kategorik değişkenleri kodlamayı içerebilir (Singh, Dwivedi ve Kumar, 2020).

Veri entegrasyonu: Bu, birden çok kaynaktan gelen verilerin analiz için kullanılabilecek tek bir veri kümesinde birleştirilmesini içerir (Swapna, Niranjan, Srinivas ve Swapna, 2016).

Veri azaltma: Bu, ilgisiz veya gereksiz verileri kaldırarak veri kümesinin boyutunu küçültmeyi içerir. Bu, analiz algoritmalarının verimliliğini artırabilir ve depolama gereksinimlerini azaltabilir (Salem ve Abdo, 2016).

Veri ayrıklaştırma: Bu, verileri bölmelere veya aralıklara bölerek sürekli verileri kategorik verilere dönüştürmeyi içerir. Bu, analiz sürecini basitleştirebilir ve sonuçların yorumlanmasını kolaylaştırabilir (Kamiran ve Calders, 2012).

Özellik seçimi: Bu, veri kümesinden en alakalı özelliklerin bir alt kümesinin seçilmesini içerir. Bu, analiz algoritmalarının performansını artırabilir ve fazla uydurma riskini azaltabilir (Misra ve Yadav, 2019).

Veri keşfi ve analizi, veri setine istatistik yöntemlerin sağladığı bakış açısının değerlendirme sonuçları ile gerçekleştirilir. Veri içerisinden elde edilen bilgiler özet istatistiki sonuçlar, grafik gösterimler ile rapor edilir. Verilerin incelenmesi sürecinde, veri setinin yapısına bakılarak öz nitelik sayısı, öz niteliklerin sürekli veya kategorik veri, veri seti içerisinde eksik değerlerin olup olmadığı, veri içerisinde belli grupları ifade eden benzersiz değerlerin olup olmadığı, veri setinde bulunan değerlerin ortalama veya medya dağılımı, öz niteliklerin birbiri ile benzerlik düzeyi incelenmektedir.

Öz niteliklerin ortaya çıkartılması sonrası hangi öz niteliğin kullanılması gerektiğinin tespiti amacıyla veri seti içerisindeki değerlerin eğilim yönünü tespit etme, öz nitelikler arası ilişkinin tespiti, yeni öz niteliklerin tespiti ve benzeri yöntemler kullanılarak öz nitelik seçimi yapılır (Celis, Keswani ve Vishnoi, 2020). Bağımlı değişkenin türüne göre algoritma seçimi gerçekleştirilmektedir. Bağımlı değişkenin sürekli veri türünde olması durumunda regresyon, kategorik olması durumunda da sınıflandırma algoritmaları tercih edilmektedir.

3.Yöntem

Türkiye’de Güneydoğu Anadolu bölgesinde bulunan bir eczaneye ait 02/01/2020 tarihi ile 07/09/2022 tarihleri arasındaki ilaç satış verileri kullanılmıştır. Çalışmada toplam veri satırı 93238 kullanılmıştır. Tablo 2’de veri seti hakkında özet bilgi verilmiştir.

Tablo 2: Verilerin Özeti

Nitelik	Bilgi
Toplam Tekil İlaç Türü	2.446
Toplam Reçete Sayısı	42.378
Satılan Toplam İlaç Sayısı	198.634
İlaç Satış Tarih Aralığı	02/01/2020 - 07/09/2022

Reçete Başına Satılan Ortalama İlaç Sayısı	4,68
Tekil Etkin Madde	708
Tekil ATC kodu	682
Tekil Genelleştirilmiş ATC Kodu	86

Veri seti incelendiğinde, ilaçların gramaj, sunum şekli değişikliğe uğramasından dolayı tekil ilaç türü sayısının yüksek olduğu anlaşılmaktadır. Bu nedenle veri seti incelenirken ilaçların etken maddeleri ve Anatomik ve Terapötik Kimyasal Sınıflandırma Sistemi (ATC) kodlarına göre ek iki veri sütunu oluşturulmuştur. Oluşturulan ilk sütun genel bir ATC sınıflandırmasını ifade ederken diğer sütun detaylandırılmış ATC kodunu ifade etmektedir (Hollingworth ve Kairuz, 2021).

ATC kodları incelendiğinde A ifadesi Sindirim sistemi ve metabolizma ilgilidir. A01AA01 kodu sodyum Florür'ü işaret etmektedir. Veri setinde 1 ve 7 basamaklı kırılım olacak şekilde iki sütun oluşturulmuştur. Bu şekilde yeni oluşturulan bağımsız değişkenler ilgili çevrede veya eczane müşterilerinde gerçekleşen genel hastalık eğilimlerini ortaya çıkarma yönüyle gereklidir (Skrbo, Begovic ve Skrbo, 2004).

Veri Setinin başarısının ölçümü için ortalama mutlak hata (mean square error) ve hata karelerinin karekökü (root mean square error) değerleri kullanılmıştır. Python dilinde ve Sklearn Kütüphanesi kullanılarak en bilinen algoritmalar kullanılarak analiz yapılmıştır. Yapılan çalışmada makina öğrenmesi algoritmaları ile bir sonraki 15 gün içerisinde yapılacak satış miktarının en yüksek başarımla tespit edilmeye çalışılmasıdır. Çalışmada makina öğrenmesi algoritmalarına ek olarak ARIMA yöntemi kullanılmıştır (Piccolo, 1990). Veri setinin özünde zaman serileri veri seti olmasından dolayı ve çalışmada amaçlanan hedefin geleceğe yönelik satış tahmini olmasından dolayı, bu alanda en uygun düşen yöntemlerden biri olan ARIMA değerlendirmeye tabi tutulmuş ve makina öğrenmesi yöntemlerine göre başarısı değerlendirilmiştir.

Veri seti aşağıdaki başlıklara göre keşifsel analizi gerçekleştirilmiştir.

- Yıl
- Ay
- Gün
- Hafta sonu
- Nöbetçi
- Hafta
- Toplam Satış Adedi
- Etkin Madde
- ATC Kodu

3.1.ARIMA

CHASE, (2013), ARIMA modelinin uygulamasını detaylı olarak açıklamaktadır. Zaman serisinin hayati bileşenlerini ve regresyon yöntemlerini bütünleştiren tahmin tekniğidir. Box, Jenkins, Reinsel ve Ljung (2015), ARIMA tahmin yöntemini tanıtmıştır ve kapsamlı bir tahmin yaklaşımı geliştirmişlerdir.

ARIMA (Otoregresif Entegre Hareketli Ortalama) modeli, tarihsel verilere dayalı tahminler yapmak için kullanılabilecek, yaygın olarak kullanılan bir zaman serisi analiz tekniğidir. Model üç bileşene dayanmaktadır: otoregresyon, fark alma ve hareketli ortalama. Otoregresyon, gelecekteki değerleri tahmin etmek için zaman serisinin geçmiş değerlerinin kullanılması anlamına gelir. Otoregresif bir

modelde, zaman serisinin mevcut değerinin geçmiş değerlerinin doğrusal bir fonksiyonu olarak tahmin edildiği bir regresyon denklemi oluşturulur. Fark alma, zaman serisini durağan hale getirmek için, yani zaman içinde sabit bir ortalama ve varyansla, ardışık değerler arasındaki farkları alma sürecini ifade eder. Doğru modelleme ve tahmin için durağanlık gereklidir. Hareketli ortalama, zaman serisinin gelecekteki değerlerini tahmin etmek için geçmiş tahmin hatalarının kullanılması anlamına gelir. Hareketli bir ortalama modelinde, zaman serisinin mevcut değerinin geçmiş tahmin hatalarının doğrusal bir fonksiyonu olarak tahmin edildiği bir regresyon denklemi oluşturulur (Piccolo, 1990).

ARIMA modeli, zaman serisinin gelecekteki değerlerini tahmin etmek için kullanılabilecek tek bir model oluşturmak için bu üç bileşeni birleştirir. Model, ARIMA(p,d,q) olarak gösterilir; burada p, otoregresif bileşenin sırası, d, zaman serisini durağan hale getirmek için gereken farkın sırası ve q, hareketli ortalama bileşeninin sırasıdır. ARIMA modeli, zaman serisi verilerini analiz etmek ve tahmin etmek için finans, ekonomi ve mühendislik dahil olmak üzere çeşitli alanlarda yaygın olarak kullanılmaktadır (Ho ve Xie, 1998; Shumway ve Stoffer, 2017).

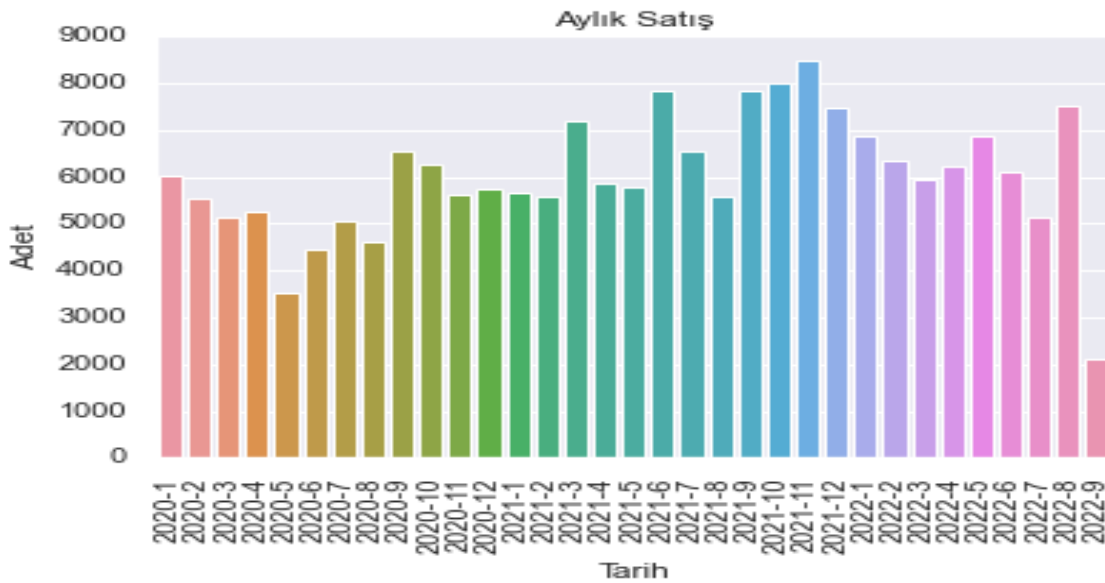
3.2. Prophet Model

Prophet, Facebook tarafından zaman serisi veri analizi için geliştirilmiş bir istatistiksel tahmin modelidir. Model, mevsimselliği, doğrusal olmayan eğilimleri ve gerçek dünya veri kümelerinde yaygın olarak bulunan zamanla ilgili diğer özellikleri işlemek için tasarlanmıştır. Prophet trend, mevsimsellik ve tatil bileşenlerinden oluşan bir katkı modeli kullanır. Trend bileşeni, verilerdeki uzun vadeli değişiklikleri yakalarken mevsimsellik bileşeni, verilerdeki periyodik dalgalanmaları yakalar. Tatil bileşeni, tatiller veya promosyonlar gibi verileri etkileyebilecek bilinen olayların etkisini modellemek için kullanılır (Shohan, Faruque ve Foo, 2022).

Prophet, model parametrelerini ve belirsizlik aralıklarını tahmin etmek için Bayes çıkarımını kullanır. Model, parçalı lineer trendler, seyrek mevsimsel etkiler ve tatil etkilerinde seyrekliği zorlamak için Laplace öncelikleri dahil olmak üzere sonsal dağılımları bilgilendirmek için çeşitli öncelikleri birleştirir. Prophet, kullanım kolaylığı ve çok çeşitli veri kümelerini işleme yeteneği nedeniyle popülerlik kazanmıştır. Model son derece özelleştirilebilir olup, kullanıcıların kendi özel kullanım durumlarına uyması için çeşitli hiperparametreler ve öncelikler belirlemesine olanak tanır. Ek olarak, Prophet, model yorumlama ve değerlendirmeye yardımcı olmak için çeşitli teşhis ve görselleştirme araçları sağlar (Jha ve Pande, 2021).

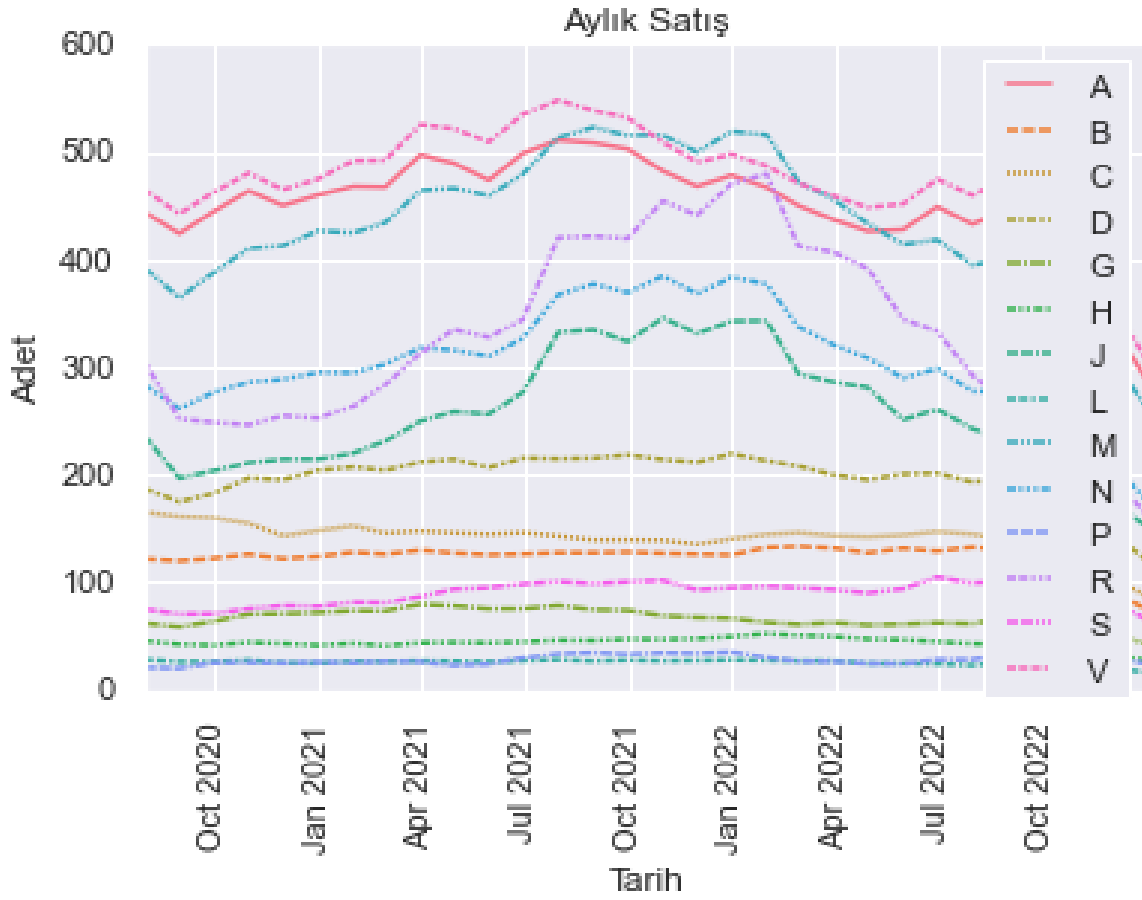
4. Bulgular

Şekil 1'de eczanenin aylık bazda yaptığı ürün satış adedi görülmektedir. Aylık satış verileri incelendiğinde en fazla satışın 8469 adet ile 2021 Aralık ayında gerçekleştiği görülmektedir.



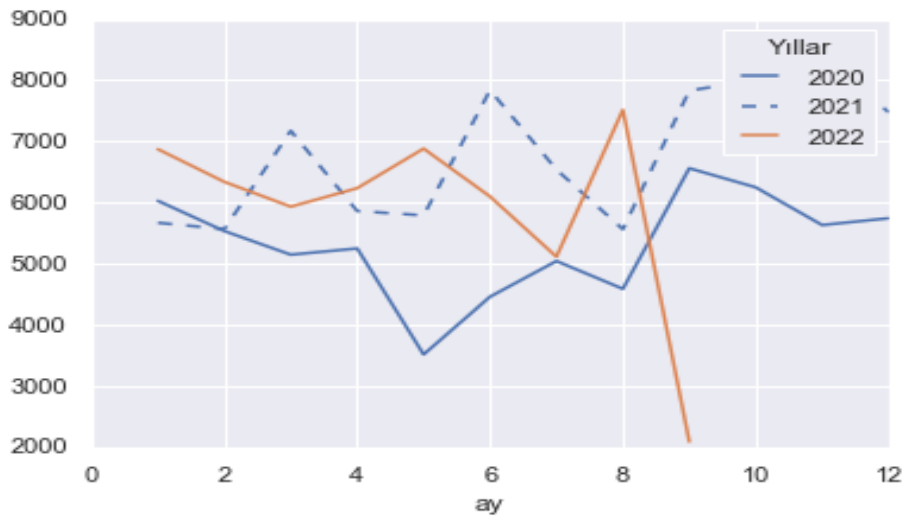
Şekil 1: Aylara Göre İlaç Satış Verileri

Satış verilerinde ATC koduna göre inceleme yapıldığında en fazla satışı yapılan türler sırası ile solunum sistemi, çeşitli ilaçlar kategorisi, sindirim sistemi ve metabolizma ilaçlarıdır. 2021 yılı sonunda ve 2022 yılı başlangıcını oluşturan dönemde enfeksiyon ve kas iskelet ilaçlarının satışının artış gösterdiği Şekil 2'den anlaşılmaktadır.



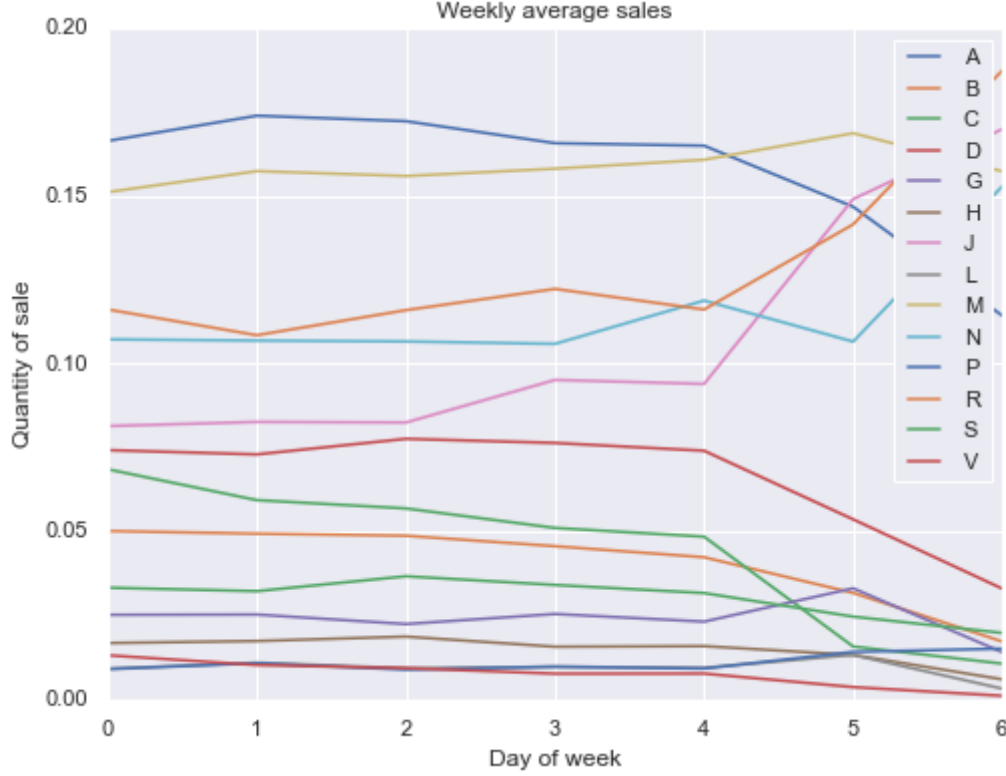
Şekil 2: Aylara Göre ATC koduna Satış Verileri

Şekil 3'te veri setinde bulunan 3 yıllık veri yıllara bölünmüş ve aylık olarak incelenmiştir. Aylık bazda incelemede yılların aynı aylarına denk gelen satış verilerinin benzerlik göstermediği ancak satılan ilaç adedinin yıllar içerisinde arttığı anlaşılmaktadır.



Şekil 3: Yılların Aynı Aylarına göre satış adedi

Şekil 4’te tüm veri seti içerisinde yapılan satışların haftanın günlerine göre ortalaması alınmış ve kas iskelet sistem ile ilgili ilaçların ortalamasının yüksek seyrettiği ve istikrarlı seyrettiği ancak enfeksiyon ve solunum ile ilgili ilaçların hafta sonunda artış gösterdiği anlaşılmaktadır.



Şekil 4: Haftanın Günlerine Göre Satış Ortalamaları

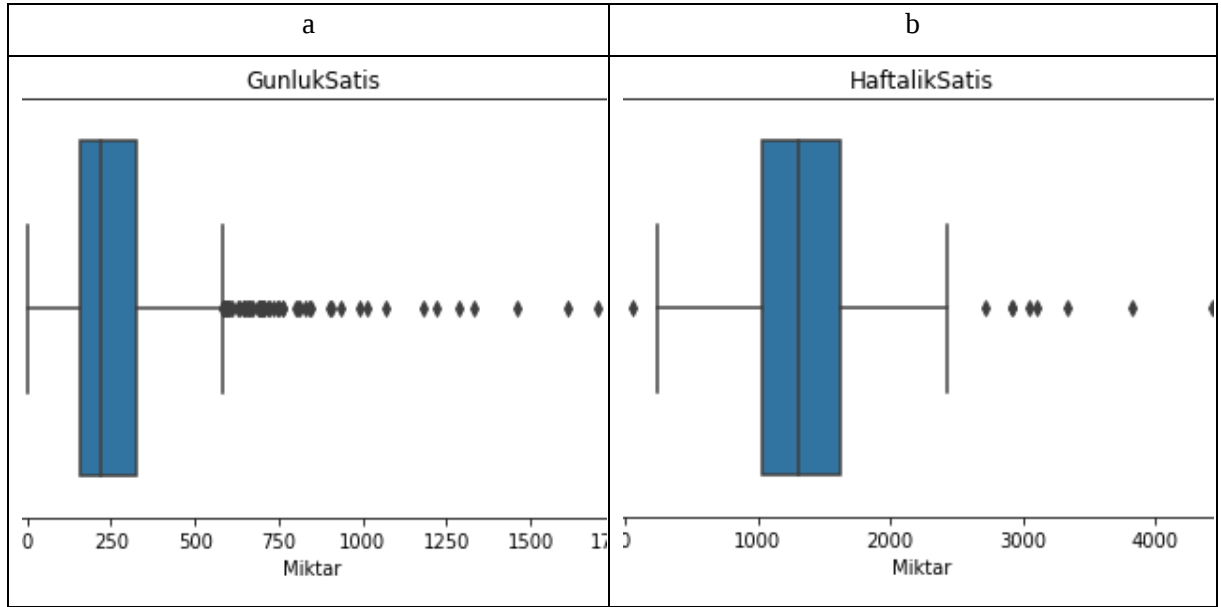
Veri seti satış verileri günlük ve haftalık olarak yeniden yapılandırılmıştır. Makine öğrenmesi algoritmalarının en bilinenleri ile yapılan çalışmada günlük ve haftalık veriler üzerinden elde edilen en başarılı sonuçlar Tablo 3’de gösterilmiştir. Veriler üzerinden elde edilen sonuçlar değerlendirildiğinde en başarılı sonuçların temel olarak rastgele orman regresresyon algoritması ve bu algoritmayı temel alarak geliştirilmiş algoritmalar olduğu görülmektedir. Makine öğrenmesi algoritmaları ile elde edilen sonuçların, günlük verilerde ortalama mutlak hata (MAE) değerinin en düşük olarak 119 bulunduğu ve korelasyon katsayısının 0.24 olduğu görülmektedir. Haftalık veri setinde en küçük MAE değeri BayesianRidge algoritması ile 492 bulunmuş olup korelasyon değeri -0.01’dir. DecisionTreeRegressor ve BayesianRidge algoritması ile elde edilen korelasyon katsayısı veri setini gerek günlük gerekse haftalıkta açıklamak için yeterli olmaktan uzaktır.

Tablo 3. Satış adetlerinin Günlük ve Haftalık Tahmin Sonuçları

	Günlük Veri			Haftalık Veri		
Model	MAE	RMSE	SCORE	MAE	RMSE	SCORE
RandomForestRegressor	139	225	0.033	644	861	-0.42
GradientBoostingRegressor	135	214	0.12	569	787	-0.19

LGBMRegressor	141	220	0.07	547	745	-0.07
OrthogonalMatchingPursuitCV	158	212	0.144	752	854	-0.40
DecisionTreeRegressor	119	199	0.24	544	730	-0.02
LarsCV	157	211	0.15	722	836	-0.34
LinearRegression	158	212	0.14	749	865	-0.44
BayesianRidge	151	208	0.17	492	727	-0.01
HuberRegressor	120	206	0.18	486	737	-0.04

Veri setinin makina öğrenmesi algoritmaları ile değerlendirildiğinde başarılı bir sonucun elde edilemediği anlaşılmaktadır. Veri setinin tarihsel bir süreç içerisinde satışları göstermesi nedeniyle bir zaman serisi problemidir. Zaman serisi problemleri için gerekli kullanılacak yöntemler değerlendirildiğinde literatürde ARIMA, Prophet algoritmalarının kullanıldığı anlaşılmaktadır.



Şekil 4: Günlük ve Haftalık Satış verilerinde Aykırı Değer Grafiği

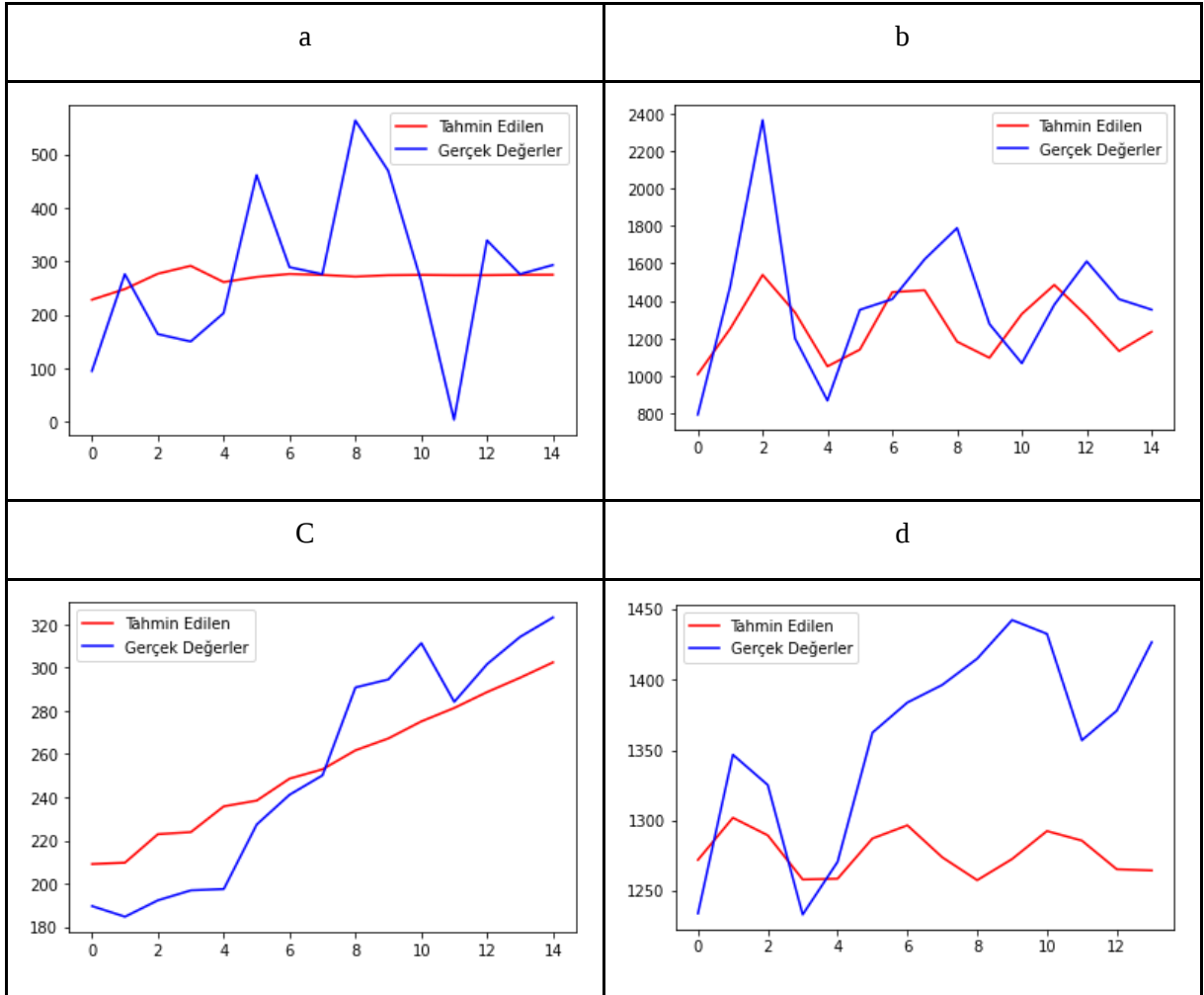
Durağan niteliğe sahip olmayan ancak fark alma işlemine tabi tutularak durağanlaştırılan serilere uygulanan modele ARIMA adı verilmektedir. ARIMA yönteminin bir diğer tanımı, durağanlaştırılmış serilere uygulanan modellere “durağan olmayan doğrusal stokastik modeller” denir (Kaynar ve Taştan, 2010). ARIMA modeli, temel olarak p, d ve q değerlerinin belirlenmesi ile hesaplanmaktadır. Model içerisinde bulunan p değeri, veri setinde bağımlı değişkenin gecikme gözlemlerinin sayısını ifade etmektedir. Veri setinde hareketli ortalama hesaplaması gerçekleştirilmektedir. Hareketli ortalamanın pencere değeri q ile ifade edilmektedir. Veri setinde, işlenen gözlem değerlerinin farklılaşma sayısını ifade etmektedir. Her veri seti için kendi niteliklerine göre p, q ve d değeri belirlenmektedir. Uygun değerlerin seçilmesi modelin başarısını oluşturmaktadır.

Veri seti hazırlanırken veri seti içerisinde bulunan aykırı değerlerin tespit edilmesi gerekmektedir. Aykırı değerler kimi zaman tahmin edilemeyen durumlar kimi zamanda hatalı veri seti girişinden

olabilmektedir. Bu nedenle tespit edilen aykırı değerler veri setinin ortalama değerleri ile değiştirilmiştir. Aykırı değerlerin grafiksel olarak Şekil 4'te görebilirsiniz. Veri setinde aykırı değerlerin işlenmesi sonrası veri seti dalgalanmaların etkisinden kurtulmak amacıyla 10 günlük ortalama alarak düzenlenmiştir. Günlük ve haftalık veriler üzerinden ve ilgili veri setlerinin ortalamasız ve ortalamalı olarak ARIMA modeli ile elde edilen sonuçlar Tablo 4'te gösterilmiş ve grafiksel gösterimi Şekil 5'te Tablo 3 ile uyumlu olarak verilmiştir.

Tablo 4: ARIMA Model Sonuçları

	Günlük		Haftalık	
Veri Seti	Parametre	RMSE	Parametre	RMSE
ortalamasız	(3, 3, 2)	139	(3, 2, 3)	323
Ortalamalı	(2, 3, 2)	23	(2, 1, 2)	103



Şekil 5: ARIMA Model sonuçları

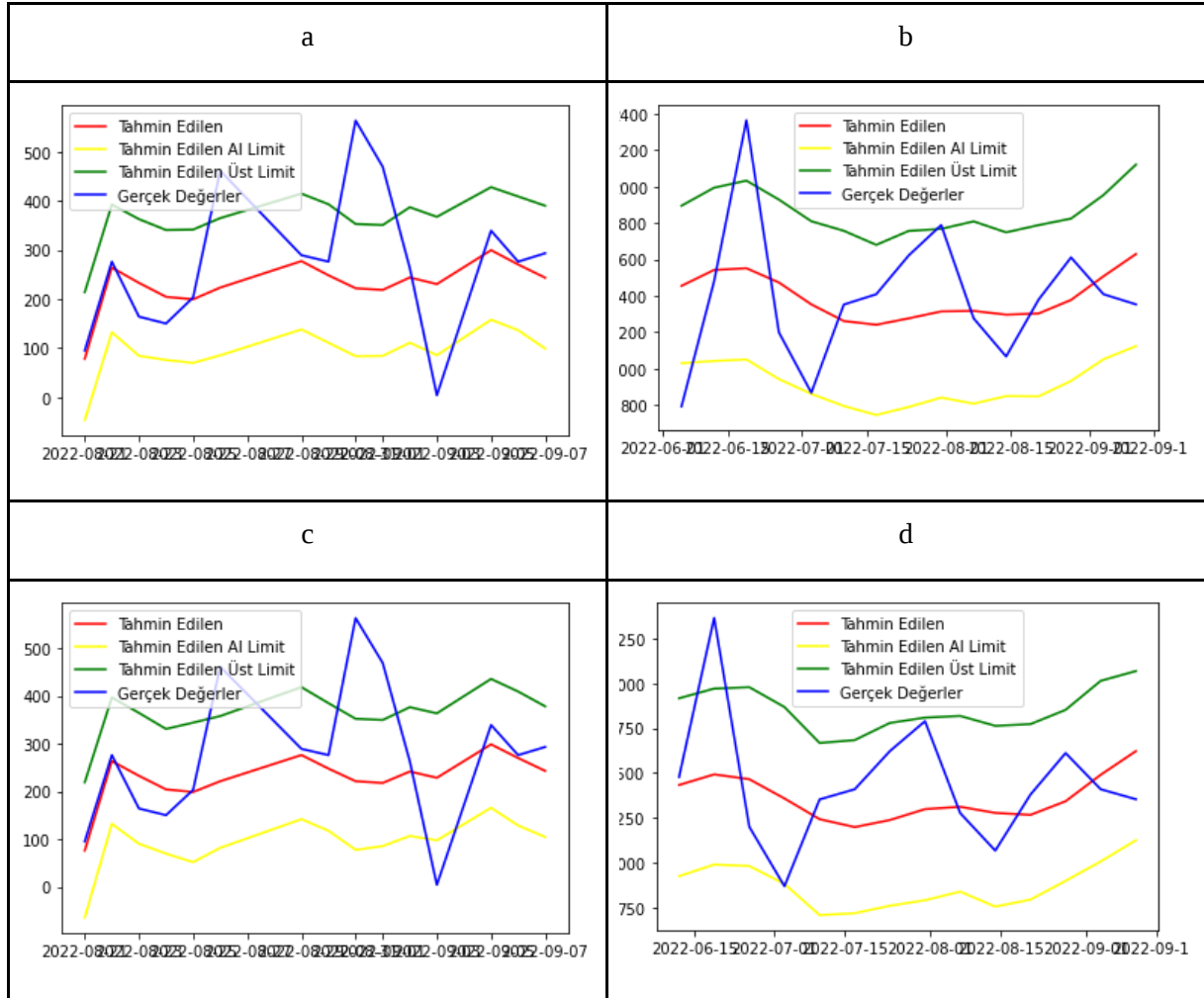
Eski ismi Facebook olan Meta firması tarafından geliştirilmiş lineer olmayan seriler için kullanılan bir algoritmadır. Dönemsel ve mevsimsel etkilere karşı dirençli yapısı ile ön plana çıkmaktadır. Prophet

algoritması veri seti içerisinde kayıpların ve uç değerlerin tolere edebilme yeteneğine sahiptir. Model, Meta firması tarafından aktif olarak kullanılmaktadır (Selvi ve Gülsoy, 2020; Taylor ve Letham, 2017).

Meta firması tarafından zaman serileri için özel olarak geliştirilmiş olan Prophet algoritması ile yapılan değerlendirmede elde edilen sonuçlar Tablo 5'te verilmiştir. Günlük sonuçlar değerlendirildiğinde ortalaması alınmış veriler ile ortalaması alınmamış veriler arasında sonuç bakımından belirgin bir farklılık olmadığı ancak haftalık verilerde az da olsa farklılık olduğu görülmüştür.

Tablo 5: Prophet Model Sonuçları

	Günlük	Haftalık
Veri Seti	RMSE	RMSE
ortalamasız	141.53	365.25
Ortalamalı	141.84	351.85



Şekil 6: Prophet Model sonuçları

5.onuç

Stok yönetimi, bir işletme için kritik karar alma süreçlerinden biridir ve bu süreçte yapılacak satışların tahmini, işletme açısından önemli bir problemdir. Bu nedenle, bu problemin çözümüne katkı sağlamak ve gelecekte muhtemel satışların tahmini için, bir eczanenin 3 yıllık satış verilerini barındıran bir veri seti üzerinde çalışma yapılmıştır. Eczane satış verileri üzerinde gerçekleştirilen çalışma ile eczanenin satışlarında var olan gizli kalmış bilgilerin ortaya çıkartılması amaçlanmış olup sonraki 15 günlük satış miktarının tahmin edilmesine yönelik çalışma gerçekleştirilmiştir. Çalışmanın küçük işletmelerin daha verimliliğinin artırılması, stok yönetiminin ve nakit akışının optimize edilebilmesi amacıyla satışlarının tahmin edilebilir seviyeye getirilebilir olması yönüyle değerli bir çalışmadır.

Eczaneler hafta sonu kapalı olan ve günün sadece belli saatlerinde açık olan işletmeler olmasının yanı sıra belirsiz aralıklar ile nöbet günlerine sahiptirler. Eczaneler nöbet günlerinde 24 saatlik satış yapmakta ve bulunduğu ilde sınırlı eczanenin satış yapmasında ötürü satış miktarı hayatın normal akışına ters olarak artış göstermektedir. Eczanenin nöbetçi olduğu günlerde gerçekleşen satış artışı veri seti içerisinde aykırı değer olarak karşımıza çıkmaktadır. Aykırı değerlerin olması veri seti üzerinde olumsuz etkiye sahip olmakta, ortalama satış miktarı ve satış verilerinde normalizasyon değerlerini olumsuz yönde etkilemektedir. Nöbet günlerinden kaynaklı gözlemlenen aykırı değerler ortalama satış verileri ile değiştirilerek, söz konusu problem kısmi olarak ortadan kaldırılmıştır.

Veri seti incelendiğinde ardışık zaman aralığına yayılmış ve sınırlı bir zaman aralığında gerçekleşen satışlar bulunmaktadır. Bu durum veri setine yönelik yaklaşımın zaman serileri bağlamında olması gerekliliğini ortaya çıkarmaktadır. Çalışmada, veri seti çeşitli makine öğrenmesi modelleri ile değerlendirmeye tabi tutulmuştur. Günlük veri üzerinde elde edilen en iyi sonuç DecisionTreeRegressor algoritması ile 119 MAE ve haftalık veri seti BayesianRidge algoritması ile 492 MAE elde edilmiştir. Elde edilen sonuçlar değerlendirildiğinde anlamlı bir sonuca varılmadığı görülmektedir. Veri seti, zaman serilerine yönelik özelleştirilmiş ve bu alanda sıklıkla kullanılan ARIMA ve Prophet algoritmaları ile gerçekleştirilen analizler sonucunda anlamlı sonuçlar elde edilmiştir.

Veri setinde zaman içerisinde gözlemlenen dalgalanmaların etkilerin azaltılması amacıyla 10 günlük zaman penceresi ile ortalama satış değerleri dikkate alınarak yapılan ve ARIMA model uygulanması sonucunda Tablo 4'te de görüleceği üzere başarılı sonuçlar elde edilmiştir. ARIMA modeli ile günlük veriler üzerinde gerçekleştirilen çalışma sonucunda 23 RMSE elde edilirken haftalık veriler üzerinde 103 RMSE sonucu elde edilmiştir. Prophet algoritması ile yapılan değerlendirmede günlük verilerin ortalama işlemine tabi tutulması durumunda 141.84 ve ortalama işlemsiz 141.53 RMSE ile tahmin başarısı gösterdiği anlaşılmaktadır. Yine haftalık veri seti ile gerçekleştirilen çalışmada ortalama değerleri alınması durumunda 351.85 iken ortalama alınmamış verilerde 365.25 RMSE elde edilmiştir. Günlük sonuçlar nazara alındığında ortalaması alınmış veriler ile ortalama işlemine tutulmamış veriler arasında sonuç bakımından kayda değer bir farklılığın olmadığı tespit edilmiştir. Haftalık verilerde ise az da olsa bir farklılık olduğu görülmüştür.

ARIMA algoritması ve satış verilerinin ortalamaya tabi tutulmadan elde edilen sonuçların Prophet algoritması ile yakın çıkması, Prophet algoritmasının uygun zaman penceresi tespiti ve buna yönelik ortalama alma işlemlerine ihtiyaç bırakmaması nedeniyle daha kullanışlı olduğu sonucuna varılmaktadır.

Zaman serilerinden oluşan veri setlerinde ARIMA algoritmasının, veri seti üzerinde uygun ön çalışmanın yapılması durumunda daha verimli olduğu ancak daha az ön işleme ihtiyaç duyarak en verimli sonuçlara yakın sonuçları elde edilebilmesi yönüyle hızlı olduğu sonucuna varılmıştır. İleride yapılacak çalışmalarda ilaçların etkin maddesi ve ATC kodlarına göre değerlendirme yapılması işletmenin daha detaylı satış öngörüsü geliştirilebileceği düşünülmektedir.

Kaynakça

- Agarwal, S. (2013, Aralık). Data mining: Data mining concepts and techniques. In *2013 international conference on machine intelligence and research advancement* (pp. 203-207). IEEE.
- Anusha, S. L., Alok, S., ve Shaik, A. (2014). Demand forecasting for the Indian pharmaceutical retail: A case study. *Journal of Supply Chain Management Systems*, 3(2), 1-8

- Aydemir, E., ve Yavuz, M. (2019). Mevsimlere Göre İlaç Satış Verilerinin Birliktelik Analizi İle İncelenmesi. *Uluslararası Yönetim Bilişim Sistemleri ve Bilgisayar Bilimleri Dergisi*, 3(1), 23-30.
- Box, G. E., Jenkins, G. M., Reinsel, G. C., ve Ljung, G. M. (2015). *Time series analysis: forecasting and control*. John Wiley & Sons.
- Bro, R., van den Berg, F., Thybo, A., Andersen, C. M., Jørgensen, B. M., ve Andersen, H. (2002). Multivariate data analysis as a tool in advanced quality monitoring in the food production chain. *Trends in Food Science & Technology*, 13(6-7), 235-244.
- Celis, L. E., Keswani, V., ve Vishnoi, N. (2020). Data preprocessing to mitigate bias: A maximum entropy based approach. In *International conference on machine learning*. PMLR. 1349-1359.
- Chase, C. W. (2013). *Demand-driven forecasting: a structured approach to forecasting*. John Wiley & Sons.
- Çetin, V., ve Yıldız, O. (2022). Veri analizinde veri ön işleme teknikleri üzerine kapsamlı bir inceleme. *Pamukkale Üniversitesi Mühendislik Bilimleri Dergisi*, 28(2), 299-312.
- Das, S. K., ve Tripathi, S. (2018). Adaptive and intelligent energy efficient routing for transparent heterogeneous ad-hoc network by fusion of game theory and linear programming. *Applied Intelligence*, 48(7), 1825-1845.
- Dey, N., Ashour, A. S., Shi, F., Fong, S. J., ve Simon Sherratt, R. (2017). Developing residential wireless sensor networks for ECG healthcare monitoring. *IEEE Transactions on Consumer Electronics*, 63(4), 442-449.
- Friedl, M. A., ve Brodley, C. E. (1997). Decision tree classification of land cover from remotely sensed data. *Remote sensing of environment*, 61(3), 399-409.
- Ghousi, R., Mehrani, S., Momeni, M., ve Anjomshooa, S. (2012, Temmuz). Application of data mining techniques in drug consumption forecasting to help pharmaceutical industry production planning. In *Proceedings of the 2012 international conference on industrial engineering and operations management* (Vol. 3, No. 6).
- Godfrey, P., Gryz, J., ve Lasek, P. (2016). Interactive visualization of large data sets. *IEEE Transactions on Knowledge and Data Engineering*, 28(8), 2142-2157.
- Hansen, M., Dubayah, R., ve DeFries, R. (1996). Classification trees: an alternative to traditional land cover classifiers. *International journal of remote sensing*, 17(5), 1075-1081.
- Ho, S. L., ve Xie, M. (1998). The use of ARIMA models for reliability forecasting and analysis. *Computers & industrial engineering*, 35(1-2), 213-216.
- Hoaglin, D. C., Mosteller, F., ve Tukey, J. W. (1983). *Understanding Robust and Exploratory Data Analysis* (Vol. 3). New York: Wiley.
- Hollingworth, S., ve Kairuz, T. (2021). Measuring medicine use: Applying ATC/DDD methodology to real-world data. *Pharmacy*, 9(1), 60.
- Jacob, L., Bohlken, J., ve Kostev, K. (2021). What have we learned in the past year? A study on pharmacy purchases of psychiatric drugs from wholesalers in the days prior to the first and second COVID-19 lockdowns in Germany. *Journal of Psychiatric Research*, 140, 346-349.
- Jerome H. Friedman. "Greedy function approximation: A gradient boosting machine.." *Ann. Statist.* 29 (5) 1189 - 1232, October 2001. <https://doi.org/10.1214/aos/1013203451>
- Jha, B. K., ve Pande, S. (2021). Time series forecasting model for supermarket sales using FB-prophet. In *2021 5th International Conference on Computing Methodologies and Communication (ICCMC)* (pp. 547-554). IEEE.
- Kalikov, A. (2006). "Veri Madenciliği ve Bir E-Ticaret Uygulaması", Yüksek Lisans Tezi, Gazi Üniversitesi, Fen Bilimleri Enstitüsü.

- Kamiran, F., ve Calders, T. (2012). Data preprocessing techniques for classification without discrimination. *Knowledge and information systems*, 33(1), 1-33.
- Kaynar, O. ve Taştan, S. (2010). Zaman Serileri Tahmininde Arıma-Mlp Melez Modeli. Atatürk Üniversitesi İktisadi ve İdari Bilimler Dergisi , 23 (3) , 141-149 . Retrieved from <https://dergipark.org.tr/tr/pub/atauniiibd/issue/2672/35003>
- Khan, M., ve Khan, S. S. (2011). Data and information visualization methods, and interactive mechanisms: A survey. *International Journal of Computer Applications*, 34(1), 1–14.
- Liu, S., Wang, X., Liu, M., ve Zhu, J. (2017). Towards better analysis of machine learning models: A visual analytics perspective. *Visual Informatics*, 1(1), 48–56.
- Meeyai, S. (2016). Logistic regression with missing data: a comparisson of handling methods, and effects of percent missing values. *Journal of Traffic and Logistics Engineering*, 4(2).
- Misra, P., ve Yadav, A. S. (2019). Impact of preprocessing methods on healthcare predictions. In *Proceedings of 2nd International Conference on Advanced Computing and Software Engineering (ICACSE)*.
- Mosteller, F., ve Tukey, J. W. (1977). Data analysis and regression. A second course in statistics. *Addison-Wesley series in behavioral science: quantitative methods*.
- Mughal, M. J. H. (2018). Data mining: Web data mining techniques, tools and algorithms: An overview. *International Journal of Advanced Computer Science and Applications*, 9(6).
- Pabinger, S., Dander, A., Fischer, M., Snajder, R., Sperk, M., Efremova, M., Krabichler, B., Speicher, M. R., Zschocke, J., ve Trajanoski, Z. (2014). A survey of tools for variant analysis of next-generation genome sequencing data. *Briefings in Bioinformatics*, 15(2), 256–278.
- Papanagnou, C. I., ve Matthews-Amune, O. (2018). Coping with demand volatility in retail pharmacies with the aid of big data exploration. *Computers & Operations Research*, 98, 343-354.
- Piccolo, D. (1990). A distance measure for classifying ARIMA models. *Journal of time series analysis*, 11(2), 153-164.
- Rautenhaus, M., Böttinger, M., Siemen, S., Hoffman, R., Kirby, R. M., Mirzargar, M., Röber, N., ve Westermann, R. (2017). Visualization in meteorology a survey of techniques and tools for data analysis tasks. *IEEE Transactions on Visualization and Computer Graphics*, 24(12), 3268–3296.
- Salem, R., ve Abdo, A. (2016). Fixing rules for data cleaning based on conditional functional dependency. *Future Computing and Informatics Journal*, 1(1-2), 10-26.
- Sevli, O., ve Gülsoy, V. G. B. (2020). Covid-19 salgınına yönelik zaman serisi verileri ile Prophet model kullanarak makine öğrenmesi temelli vaka tahminlemesi. *Avrupa Bilim ve Teknoloji Dergisi*, (19), 827-835.
- Shohan, M. J. A., Faruque, M. O., & Foo, S. Y. (2022). Forecasting of electric load using a hybrid LSTM-neural prophet model. *Energies*, 15(6), 2158.
- Shumway, R.H., Stoffer, D.S. (2017). ARIMA Models. In: *Time Series Analysis and Its Applications*. Springer Texts in Statistics. Springer, Cham. https://doi.org/10.1007/978-3-319-52452-8_3
- Siddiqui, R., Azmat, M., Ahmed, S., ve Kummer, S. (2022, Nisan). A hybrid demand forecasting model for greater forecasting accuracy: the case of the pharmaceutical industry. In *Supply Chain Forum: An International Journal* (Vol. 23, No. 2, pp. 124-134). Taylor & Francis.
- Singh, S. K., Dwivedi, D., ve Kumar, R. (2020). Data mining: dirty data and data cleaning. *International conference on Recent Trends in Artificial Intelligence, IOT, Smart Cities & Applications*

- Skrbo, A., Begović, B., ve Skrbo, S. (2004). Classification of drugs using the ATC system (Anatomic, Therapeutic, Chemical Classification) and the latest changes. *Medicinski arhiv*, 58(1 Suppl 2), 138-141.
- Soley-Bori, M. (2013). Dealing with missing data: Key assumptions and methods for applied analysis. *Boston University*, 4(1), 1-19.
- Swapna, S., Niranjana, P., Srinivas, B., ve Swapna, R. (2016, Mart). Data cleaning for data quality. In *2016 3rd International Conference on Computing for Sustainable Global Development (INDIACom)* (pp. 344-348). IEEE.
- Taylor SJ, Letham B. (2017). Forecasting at scale. *PeerJ Preprints* 5:e3190v2
- Tufféry, S. (2011). *Data mining and statistics for decision making*. John Wiley ve Sons.
- Tukey, J. W. (1977). *Exploratory data analysis* (Vol. 2, pp. 131-160). Pearson doi: https://doi.org/10.1007/978-3-662-45006-2_9
- Van Boeckel, T. P., Gandra, S., Ashok, A., Caudron, Q., Grenfell, B. T., Levin, S. A., ve Laxminarayan, R. (2014). Global antibiotic consumption 2000 to 2010: an analysis of national pharmaceutical sales data. *The Lancet infectious diseases*, 14(8), 742-750.
- Zhang, Z. (2016). Missing data imputation: focusing on single imputation. *Annals of translational medicine*, 4(1).

Research Article**Keşifçi Veri Analizi ile Eczane Satış Analizi ve Satış Tahmini***Pharmacy Sales Analysis and Sales Forecasting with Exploratory Data Analysis***Mansur BEŞTAŞ**

Dr. Siirt, Türkiye

mansur@siirt.edu.tr<https://orcid.org/0000-0002-8192-2044>**Extensive Summary**

The consumption of drugs has increased alongside their increased accessibility and daily use, given their importance in human health. In the pharmaceutical industry, it is crucial to understand pharmacy sales, where the industry has direct contact with end-users, and to develop forward-looking strategies. A considerable share of pharmaceutical industry sales is derived from individual purchases of medication. Unlike consumer goods, health products are primarily consumed as a necessity, and are less influenced by socio-economic factors and promotional activities. Exploratory analysis of health sector consumption can help identify changes in public health in a given area. Accurately estimating future sales is critical to optimizing enterprise efficiency. In this study, exploratory data analysis was conducted on three years of sales data from a pharmacy located in the Southeastern Anatolia region, and machine learning and time series algorithms were applied to forecast sales for the next 15 days. The ARIMA method produced the most successful results, with a root mean square error (RMSE) of 23.

Inventory management is a crucial aspect of decision-making for businesses, and forecasting future sales is an essential problem that businesses must address. This study aims to contribute to the solution of this problem by analyzing a dataset containing three years of sales data from a pharmacy, to predict possible sales in the future. The study aims to reveal confidential information in the pharmacy's sales and estimate the sales amount for the next 15 days. This study is valuable because predicting sales can increase the efficiency of small businesses, optimize stock management and cash flow. Pharmacies are businesses that have specific opening hours and are closed on weekends. Due to the limited number of pharmacies in the province, sales increase during the pharmacy's on-duty hours, which appears as an outlier in the data set. Outliers negatively affect the data set and the average sales amount, and normalization values in the sales data. The problem of outliers has been partially eliminated by replacing observed outliers with average sales data.

The data set contains sales spread over consecutive time intervals, realized in a limited time interval, indicating that the approach to the data set should be in the context of time series. The study evaluated the data set using various machine learning models, including DecisionTreeRegressor, BayesianRidge, ARIMA, and Prophet algorithms. The study obtained the best result with the ARIMA and Prophet algorithms, which are customized for time series and frequently used in this field. Successful results were obtained as a result of the application of the ARIMA model, which considered the average sales values with a ten-day time window to reduce the effects of the fluctuations observed in the data set over time. The study showed that the ARIMA algorithm is more efficient in datasets consisting of time series, with appropriate preliminary work, but it requires less preprocessing than the Prophet algorithm.

Future studies can develop a more detailed sales forecast of the company by evaluating the active substance of the drugs and ATC codes. Inventory management is an essential aspect of decision-making for businesses as it involves managing the supply and demand of products. Accurate forecasting of sales is crucial to ensure that businesses can maintain optimal stock levels while minimizing costs. However,

forecasting sales accurately can be a challenging task, particularly for small businesses. For this reason, this study aims to contribute to the solution of this problem by analyzing a dataset containing three years of sales data from a pharmacy to predict possible sales in the future. The study aimed to reveal confidential information in the pharmacy's sales and estimate the sales amount for the next 15 days. This study is valuable because predicting sales can increase the efficiency of small businesses, optimize stock management and cash flow. The analysis of the sales data can help businesses to make informed decisions, such as which products to stock and when to order new inventory. Additionally, predicting sales can help businesses to identify patterns and trends in consumer behavior, allowing them to adjust their strategies to meet customer demands.

Pharmacies are businesses that have specific opening hours and are closed on weekends. Due to the limited number of pharmacies in the province, sales increase during the pharmacy's on-duty hours, which appears as an outlier in the data set. Outliers are data points that are significantly different from the other data points in the dataset. Outliers can negatively affect the data set and the average sales amount and normalization values in the sales data. Outliers can also skew statistical analyses and result in incorrect conclusions. The problem of outliers has been partially eliminated by replacing observed outliers with average sales data. The data set contains sales spread over consecutive time intervals, realized in a limited time interval, indicating that the approach to the data set should be in the context of time series. Time series data is a collection of observations collected over time, such as sales data, stock prices, or weather data.

Additionally, the study has highlighted the importance of time series analysis for inventory management in small businesses such as pharmacies. The results of this study can provide insights for small business owners on how to optimize their stock management and cash flow by predicting future sales more accurately. The findings can also contribute to the development of more effective inventory management systems for small businesses. Furthermore, the study has addressed the issue of outliers in the sales data, which can negatively affect the accuracy of the sales forecasting. The presence of outliers can be problematic as they can skew the average sales data and normalization values, leading to inaccurate predictions. The study has attempted to mitigate this problem by replacing the observed outliers with the average sales data. The data set used in this study consisted of three years of sales data from a pharmacy. The sales were spread over consecutive time intervals and realized in a limited time interval, indicating that the data set should be analyzed in the context of time series. The study evaluated the data set using various machine learning models, including DecisionTreeRegressor, BayesianRidge, ARIMA, and Prophet. The best results were obtained with the ARIMA and Prophet algorithms, which are specialized for time series analysis.

The study found that the ARIMA algorithm was more efficient in datasets consisting of time series, provided that appropriate preliminary work was done on the data set. However, the Prophet algorithm was found to be more useful as it does not require appropriate time window detection and averaging processes. The results obtained without averaging the ARIMA algorithm and sales data were found to be close to the Prophet algorithm. Therefore, the Prophet algorithm is recommended as a more efficient and effective method for forecasting sales in small businesses such as pharmacies. In future studies, it is suggested that a more detailed sales forecast of the company can be developed by evaluating the active substance of the drugs and ATC codes. This can provide additional insights for small business owners on how to optimize their inventory management and improve their sales forecasting accuracy.

Inventory management is a critical decision-making process for businesses, and accurate sales forecasting is essential for optimizing stock management and cash flow. The study has demonstrated the importance of time series analysis for sales forecasting in small businesses such as pharmacies. The findings can provide insights for small business owners on how to improve their inventory management systems and increase the efficiency of their operations. The study has also highlighted the need for appropriate data preprocessing to mitigate the effects of outliers and ensure accurate predictions. The use of machine learning algorithms such as ARIMA and Prophet can be an effective method for forecasting sales in small businesses. Finally, future studies can explore more detailed sales forecasts by evaluating the active substance of the drugs and ATC codes.