



Online Gıda Perakende Sektöründe Talep Tahmin Analizi

Demand Prediction Analysis in the Online Food Retail Sector

Gülfem Işıklar Alptekin^{1*}, Utku Bozdoğan²

¹ Galatasaray Üniversitesi, Bilgisayar Mühendisliği Bölümü, gisiklar@gsu.edu.tr
ORCID: <https://orcid.org/0000-0003-0146-1581>

² Vanora Yazılım Teknolojileri AŞ, utku.bozdogan@vanoraventures.com
ORCID: <https://orcid.org/0000-0002-8912-9446>

MAKALE BİLGİLERİ

Makale Geçmişi:

Geliş 18 Ekim 2024
Revizyon 13 Şubat 2025
Kabul 12 Nisan 2025
Online 30 Haziran 2025

Anahtar Kelimeler:

*Talep tahmini, online gıda
sektörü, zaman serisi analizi,
Uzun-Kısa Süreli Bellek (LSTM)*

ÖZ

Gıda talebinin tahmin edilmesi, özellikle e-ticaret alanında iş süreçlerinin optimizasyonu ve sürdürülebilir kalkınma için büyük önem taşımaktadır. Bu çalışmada, öğrenme tabanlı algoritmalar ve zaman serisi analizlerini kullanarak geliştirilen bir talep tahmin sistemi sunulmuştur. Bu sistem, müşteri taleplerini daha doğru bir şekilde öngörmeyi ve perakende şirketlerine stratejik öneriler sunmayı amaçlamaktadır. Yapılan sayısal uygulamalar, Türkiye'deki bir gıda perakende şirketinin son yedi yıllık verileri üzerinde gerçekleştirilmiştir. Çalışmada, özellikle izolasyon ormanı, aglomeratif ve spektral kümeleme, Uzun Kısa Süreli Bellek ve Ekstrem Gradyan Arttırma gibi yöntemler kullanılmıştır. Geliştirilen modeller, satış verilerinin doğru tahmin edilmesine ve stok yönetimi gibi kritik iş süreçlerine önemli katkılar sağlamaktadır. Sonuçlar, özellikle pandemi sürecinde ve mevsimsel etkilerin dikkate alındığı durumlarda gıda talebinin daha etkin bir şekilde yönetilmesine olanak tanımaktadır.

ARTICLE INFO

Article history:

Received 18 October 2024
Received in revised form 13 February 2025
Accepted 12 April 2025
Available online 30 June 2025

Keywords:

*Demand prediction, online food sector,
time series analysis, Long Short-Term
Memory (LSTM)*

ABSTRACT

Predicting food demand is of great importance, especially for optimizing business processes and achieving sustainable development in the e-commerce sector. This study presents a demand forecasting system developed using learning-based algorithms and time series analysis. The system aims to accurately predict customer demands and provide strategic recommendations to retail companies. Numerical applications have been conducted using data from the past seven years of a food retail company in Turkey. Specifically, methods such as isolation forest, agglomerative and spectral clustering, Long Short-Term Memory (LSTM), and Extreme Gradient Boosting (XGBoost) have been utilized. The developed models significantly contribute to accurate prediction of sales data and critical business processes like inventory management. The results facilitate more effective management of food demand, particularly during the pandemic and in scenarios where seasonal effects are considered.

DOI: 10.24012/dumf.1569799

* Sorumlu Yazar

Giriş

Gıda talep tahmini hem işletmeler hem de sürdürülebilir kalkınma açısından kritik bir meseledir. Küresel gıda ve içecek e-ticaret pazarının büyüklüğü, 2022'de 57,09 milyar \$'dan 2023'te %22,2'lik bir yıllık büyüme oranı ile 69,77 milyar \$'a çıkmış, 2027'de ise %19,1'lik bir yıllık büyüme oranı ile 140,42 milyar \$'a çıkması beklenmektedir (Food And Beverages ECommerce Global Market Report 2023).

Türkiye'de ise 2019 – 2022 yılları arasında ülkenin e-ticaret hacmi yüzde 80,6 oranında oldukça yüksek bir yıllık bileşik büyüme kaydetmiştir. Euromonitor tahminlerine göre Türkiye'deki e-ticaret hacmi, 2020- 2025 yılları arasında yüzde 23,7 yıllık bileşik büyüme oranı ile 2025'te mevcut hacminin 2,3 katı bir büyüklüğe ulaşacaktır.

PwC araştırması “Türkiye'deki KOBİ'lerin E-Ticaret Görünümü” raporunda 2022 yılında Türkiye'de gıda ve süpermarket e-ticaret hacmi toplamda %3.3'lük bir pazar dilimine sahiptir. Globale bakıldığında bu oran %28 olarak görülmektedir. Türkiye'nin gıda ve süpermarket e-ticareti açısından büyüme açısından hala doyumluğa ulaşacak alanı bulunmaktadır. Günümüzde veri analizi, iş süreçlerini sürekli iyileştirmek ve stratejik kararlar almak adına önemli bir araç haline gelmiştir. Özellikle perakende sektöründe satış verilerinin doğru bir şekilde analiz edilmesi, taleplerin öngörülmesi ve stok yönetimi gibi süreçlerin etkinliği açısından kritik bir rol oynamaktadır.

Bu makalede, e-ticaret perakende şirketlerinin karşılaştığı operasyonel zorluklara yönelik kullanılabilecek bir talep tahmin sistemi geliştirilmektedir. Öğrenme esaslı algoritmalar ve zaman serisi analizini esas alan bu yazılımın temel amacı, müşteri taleplerini daha hassas bir şekilde tahmin etmek ve stratejik öneriler sunmaktır. Sayısal uygulamalar, gıda perakende sektöründe çalışan özel bir şirketin son yedi senelik verileri ile gerçekleştirilmiştir. Bu çalışmada gıda verileri özelinde çalışılsa da yazılımın kapsamı genişleterek, e-ticarete faaliyet gösteren farklı sektörler için de rahatça uyarlanabilir. Seçilen yöntemler, özellikle son yıllarda online gıda verisi üzerinde çalışan yayınlara odaklanarak yapılan akademik literatür ve dergi incelemesi sonucu belirlenmiştir.

Bu çalışma, Ocak 2016 ile Eylül 2023 tarihleri arasındaki gıda satış verilerini inceleyerek, satış trendlerini analiz etmeyi ve geleceğe yönelik satış miktarlarının tahmin edilebilmesi için uygun modellemeleri geliştirmeyi amaçlamaktadır. Veri kümesi, kişisel verilerden arındırılmış ve anonimleştirilmiş bir yapıdadır. Veriyi daha iyi anlamak amacıyla temel istatistiki analizler gerçekleştirilmiş, siparişlerin tarihsel dağılımları, ürün gruplarının satış performansları ve satış miktarlarındaki dönemsel değişiklikler ayrıntılı bir şekilde incelenmiştir. Özellikle pandemi sürecinin satışlar üzerindeki etkisi ve farklı ürün gruplarının mevsimsel eğilimleri, çalışmada dikkat çeken başlıca unsurlar arasında yer almaktadır.

Verinin işlenmesi sırasında, eksik ve mükerrer kayıtların giderilmesi için ön işleme adımları uygulanmış, bu sayede temiz ve analiz edilmeye uygun bir veri kümesi elde edilmiştir. Ayrıca, taze gıda ürünleri gibi hızla bozulabilen ürünlerin satışlarına yönelik özel bir ilgi gösterilmiştir. Çalışmanın temel amacı, bu veri kümesi üzerinde çeşitli zaman serisi analizleri ve makine öğrenimi teknikleri

kullanarak, satışların doğru bir şekilde öngörülebilmesi ve stok yönetimi gibi operasyonel süreçlere katkı sağlayacak modeller geliştirmektir. Makalede, anomali tespiti için izolasyon ormanı (*isolation forest*), veriyi kümelere ayırırken boyut düşürme amaçlı ana bileşen analizi (PCA-*Principal Component Analysis*), TSNE, UMAP ve kümeleme amacıyla *Kmeans*, Gauss karışım modelleri (*Gaussian mixture models*), DBSCAN, aglomeratif ve spektral kümeleme algoritmaları uygulanmıştır. Sonrasında tahmin için Uzun Kısa Süreli Bellek (LSTM - *Long Short-Term Memory*) ve Ekstrem Gradyan Arttırma (XGBoost - *Extreme Gradient Boosting*) modellerinden faydalanılmıştır.

Makalenin bir sonraki bölümünde, akademik literatürdeki benzer çalışmalardan bahsettikten sonra, kullanılan yöntemler kısaca tanıtılacaktır. Ardından ayrıntılı şekilde veri ön analiz, işlenmesi ve sonuçlar verilerek, en son bölümde elde edilen çıktılar tartışılacaktır.

Literatür Taraması

Online gıda perakende sektöründe öğrenme tabanlı modellerin kullanılması, gıda talebinin tahmin edilmesi açısından kritik öneme sahiptir. Bu modeller, sürdürülebilir kalkınmayı ve iş süreçlerinin optimizasyonunu destekler. Talep tahmini konusunda akademik literatürde son derece fazla çalışma olsa da konu ile en yakından ilgili çalışmalar incelendiğinde, aşağıdaki örneklerle karşılaşılmıştır:

Polonya'da gerçekleştirilen bir araştırma, gıda talebini tahmin etmek için doğrusal olmayan otoregresif dışsal sinir ağı (NARXNN) kullanmıştır [1]. Bu çalışmada, gıda atıklarını azaltmayı ve geleneksel RNN'lerden daha iyi tahminler sağlayan NARXNN'nin avantajlarını kullanmayı hedeflemişlerdir. Çalışmada kullanılan eğitim veri kümesi, sinir ağına eğitim sırasında sunulan talep değerlerini içeren bir zaman serisinden oluşmuştur. Model performansı, MAE (*Mean Absolute Error*), MAPE (*Mean Absolute Percentage Error*) ve RMSE (*Root Mean Square Error*)'nin en düşük değerleri ile belirlenme katsayısı (R-kare) en yüksek değeri esas alınarak analiz edilmiştir. Çalışmanın bulgularına dayanarak, yapay zekâ tabanlı NARX modelleri kullanılarak çok doğru gıda talebi tahminlerinin yapılabileceği sonucuna varılmıştır.

Bir diğer çalışma, Genpact tarafından yayınlanan '*Food Demand Forecasting*' veri kümesini kullanarak sipariş sayılarını tahmin etmeyi amaçlamıştır [2]. Bu çalışmada, talebi etkileyen çeşitli faktörler analiz edilmiş, etkili özellikler çıkarılmış ve yedi farklı regresör karşılaştırılmıştır. Bu regresörler arasında, rastgele orman regresörü, *Cat Boost* regresör, aşırı radyan artırma regresörü, gradyan artırma regresörü, hafif gradyan artırma makinesi regresörü, uzun kısa süreli bellek ve iki yönlü LSTM regresörleri kullanılmıştır. MAE ve RMSE değerlerine bakıldığında, rastgele orman regresörü ile doğru bir tahmin yapılabileceğini kanıtlamışlardır.

Türkiye'de yapılan bir araştırma, öğle yemeği gibi belirli zaman dilimlerinde ön rezervasyon yapılmadan yoğun gıda talebini tahmin etmek üzere makine öğrenimi tabanlı tahmin modelleri kullanmış ve bu modeller arasında yapay sinir ağları, Gauss süreci regresyonları, destek vektör

regresyonları ve karar ağaçları yer almıştır [3]. Çalışma, yöntem çeşitliliği, tahmin performansı ve uygulama kapsamı bakımından önemli bir katkı sunmuştur. En iyi performans gösteren model, mükemmel tahmin metrikleri sağlayan *ensemble* karar ağacı modeli olmuştur; MSE (*Mean Square Error*) 0.51, MAE 0.50 ve R-kare değeri 0,96 olarak ölçülmüştür (Acı ve Yergök, 2023).

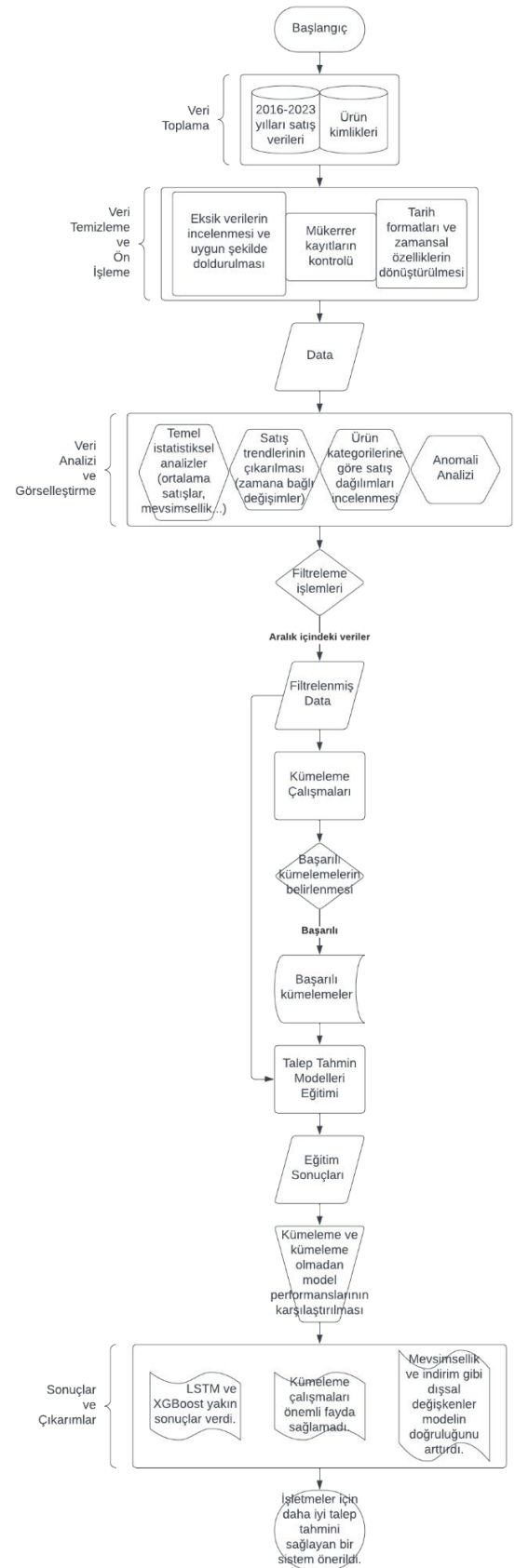
Kullanılan Yöntemler

Bu çalışmada başvuru üç temel yöntem aşağıda detaylandırılmıştır:

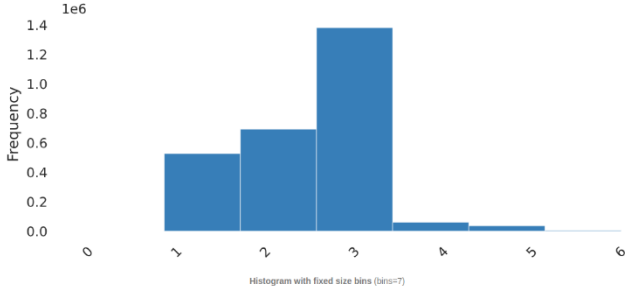
İzolasyon Ormanı (Isolation Forest) [4]: Özellikle anormal tespiti için tasarlanmış olan bu algoritma, veri kümesindeki anormal verilerin sayıca az ve özellikleri itibarıyla farklı olmasından yararlanır. Anormal veriler, normal verilere kıyasla daha kolay izole edilebilir. Algoritma, rastgele bir özellik seçer ve bu özelliğin değer aralığı içinde rastgele bir bölme noktası belirler. Bu bölme işlemi, her düğümün bir özellik üzerinde bir bölünmeyi temsil ettiği bir ağaç yapısı oluşturarak yinlemeli olarak devam eder. Anormal veriler, farklılıkları nedeniyle genellikle ağacın daha erken dallarında izole edilir, bu da onların kökten yaprağa olan mesafesinin daha kısa olduğu anlamına gelir. Aksine, normal veriler daha yaygın ve birbirine benzer oldukları için izole edilmeleri daha fazla bölünme (daha uzun yollar) gerektirir. Ormandaki ağaçların çoğunun kısa yollarla ulaştığı veri noktaları bu şekilde anormal kabul edilir. Bu yöntem, tek ve çok değişkenli veriler üzerinde etkili olup, modelin beklenen anormallik oranını da ayarlayabilme imkânı sunar.

Aglomeratif kümeleme: Hiyerarşik kümelemenin bir türü olan aglomeratif kümeleme, her gözlemi ayrı bir küme olarak kabul ederek başlar ve benzer kümeleri birleştirerek ilerler [5]. Bu yöntem, ‘aşağıdan yukarıya’ yaklaşımı benimser; yani her gözlem önce kendi başına bir küme olarak değerlendirilir, ardından en benzer iki küme birleştirilir. Bu süreç, tek bir kümeye ulaşana kadar devam eder. Kümelerin birleştirilmesi sırasında farklı bağlantı (*linkage*) yöntemleri ve uzaklık ölçütleri kullanılabilir. Aglomeratif kümeleme, önceden belirlenmiş bir küme sayısını gerektirmez ve karmaşık hiyerarşik yapıları başarıyla ele alabilir. Önceden küme sayısının belirtilmesini gerektirmediği ve kendi içinde hiyerarşik yapılarda başarılı olduğu için tercih edilmiştir.

Spektral kümeleme [6]: Veri kümesindeki benzerlikler üzerinden oluşturulan matrislerin özdeğer (*eigenvalue*) ve özvektörlerini (*eigenvector*) kullanan güçlü ve esnek bir kümeleme yöntemidir. Geleneksel kümeleme yöntemlerinin başarısız olduğu karmaşık geometrilere sahip veri kümelerini etkin bir şekilde ayırt edebilir. Spektral kümeleme, verileri özdeğer bilgisini kullanarak daha düşük boyutlu bir uzaya indirir ve ardından bu yeni temsilde K-means gibi bir kümeleme algoritması uygular. Bu yöntem, özellikle geleneksel metotlarla zorlanılan veri kümelerinde daha belirgin kümeleme sonuçları elde etmek için kullanılır. Çalışmada kullanılan yöntemler ve genel akış, Şekil 1’de özetlenmiştir.



Şekil 1. Kullanılan metodoloji ve akışı



Şekil 6. *Toplama Türü* uyarınca ürünlerin satış dağılımları

Şekil 6'da *Toplama Türü* bilgilerine bakılarak, ürünlerin satışlarının dağılımlarına bakılmıştır. Toplama türüne göre ürünlerin satış dağılımının dengeli olmadığı görülmüştür. Toplama türündeki değerlerin ayrıntılı açıklaması şu şekilde verilebilir:

- 6, tartılan balık ürünleri (ör: levrek, çipura)
- 5, tartılan ve çoklu satılmayan ürünler (ör: kırmızı lahana, beyaz lahana, tavuk bonfile, bal, gouda peyniri)
- 4, tartılan ve adetle satılan ürünler (ör:kavun, karpuz, paket kiraz, paket piliç)
- 3, tartılan ve çoklu satılan ürünler (ör:patlıcan, salatalık, kuru soğan)
- 2, tartılmayan ve adetle satılan ürünler (ör:karnabahar, kıvırcık, yufka, simit)
- 1, paketli ürünler (ör:organik domates kg, ceviz içi, kaju fıstık, hindistan cevizi sütü)
- 0, test ürünü olup toplamda 3 kayıt bulundurmaktadır.

Talep tahmini yaparken ana hedef, çabuk bozulmaları dolayısıyla zararı önlemek için başarılı talep tahmini gerektiren taze gıda ürünlerini incelemektir. Bu ürünler de çoğunlukla 3 numaralı toplama türüne ait olmakla beraber, diğer türlerde de ilgili ürünler bulunduğu görülmektedir.

Eksik ve Mükerrer Kayıtlar

Veri kümesinde yalnızca 43 adet mükerrer kayıt bulunduğu, bazı ürünlerde birden çok sipariş verildiğinde kaydın bu şekilde tutulduğu tespit edilmiş, işlem yapılmamıştır.

Veri kümesindeki alanlardan yalnızca:

1. *Tahmini Ağırlık* değerlerinin %64,7 oranında eksik olduğu belirlendi. Bu sebeple, bu değer sadece filtreleme amacıyla kullanıldı.
2. *Vergili Ürün Fiyatı* değerlerinin 29 Temmuz 2019 tarihinden itibaren saklandığı, dolayısıyla %26,2 oranında eksik olduğu belirlendi. Bu sebeple, onun yerine eksiksiz olan *Ürün Fiyatı* kullanılması tercih edildi.

Bazı alanların aldığı değerlere göre, şöyle anlamlar/kurallar çıkarılabileceği tespit edildi:

- *Sipariş Numarası* 0 değilken *Seçilen Miktarı*'nın 0 olması ürün stoğunun bittiğine işaret eder.
- *Gerçek Ağırlık* tam olarak 1 ve *Tahmini Ağırlık* eksik ise, ürün taze gıda ürünü değildir.

- Aynı siparişte veya aynı gün içerisinde bir ürünün birden fazla *Ürün Fiyatı* ile bulunması o üründe indirim işaret eder.



Şekil 7. Verideki sayısal alanlara ait korelasyon ısı haritası

Verideki sayısal alanların korelasyon ısı haritası Şekil 7'de verilmiştir. Ciddi korelasyonlar, çoğunlukla zaten birbirleriyle ilişkili olan alanlarda görülmüştür: *Ürün Fiyatı* (*UnitPrice*) ve *Vergili Ürün Fiyatı* (*UnitPriceWithTax*), *Sipariş Miktarı* (*OrderAmount*) ve *Seçilen Miktar* (*SelectedAmount*) gibi. *Sipariş Numarası* (*OriginalId*) ile *Ürün Fiyatı* ve *Vergili Ürün Fiyatı* alanlarının korelasyonu, zaman içerisinde hem verilen siparişler dolayısıyla sipariş numarasının artması hem de ürün fiyatlarının artmasıyla açıklanabilir. Satış ile ilgili veride, 3346 farklı *Gıda Kimliği* (*FoodId*) mevcuttur.

Ürün Verileri

Ürünlerin *Gıda Kimliği* bilgisine haritalandırılabilen bir alan üzerinden, ürünlere ait 'Kategori' ve 'Alt Kategori' bilgilerinin bulunduğu başka bir veri kümesine erişilmiştir. Bu veri kümesinde, 5287 farklı *Gıda Kimliği* mevcuttur.

Bu veri kümesinde 12 farklı kategori ve 54 farklı alt kategori bulunmaktadır (Şekil 8). Kategoriler ve ilgili alt kategoriler aşağıdaki gibi listelenebilir:

Ekmek ve Unlu Mamüller: Ekmek - Sos - Yufka - Makarna, Mantı & Un - Çikolata & Tatlı - Kuru Gıdalar
Et, Tavuk, Balık: Et & Et Ürünleri: Tavuk - Balık & Deniz Ürünleri - Sos

Kahvaltılık: Peynir - Yumurta - Reçel - Yoğurt - Süt Ürünleri - İçecekler - Ekmek - Organik Kahvaltılık - Ezme

Et & Et Ürünleri: Sebzeler - Pekmez,Tahin & Sirke - Kahvaltılık - Zeytinyağı - Sos - Granola - Çikolata & Tatlı

Kuru Gıdalar: Kuruyemiş - Çay & Kahve - Baharat & Tuz - Bakliyat - Erişte & Tarhana - Makarna, Mantı & Un - Ezme - Çikolata & Tatlı - Sos - Glutensiz Ürünler - İçecekler - Pekmez,Tahin & Sirke - Anne & Çocuk - Ev Yemekleri - Kuru Gıdalar - Granola - Ekmek - Süt Ürünleri

Light Ürünler: Light Ürünler - Ekmek - Granola

Meyveler: Meyveler - Reçel

Organik Ürünler: Organik Sebzeler- Organik Kahvaltılık - Organik Et, Tavuk - Organik Bal, Pekmez & Tahin - Ezme - Organik Çay & Kahve - Organik Zeytin & Zeytinyağı - Organik Bakliyatlar - Organik Makarna, Mantı, Un - Organik Salça, Sirke & Soslar - Organik Kuruyemişler

- Organik Baharat & Tuz - Organik Meyveler - Yoğurt - Organik Çikolata & Tatlı - İçecekler - Granola - Çikolata & Tatlı

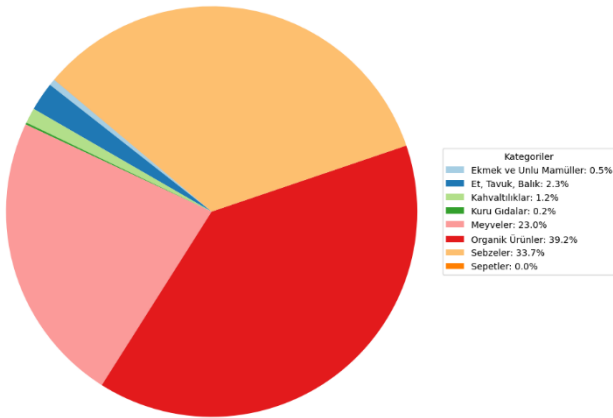
Salça, Sirke & Soslar: Salça: Sirke - Sos - Turşu - İçecekler - Ezme - Sebzeler

Sebzeler: Sebzeler - Sos - Organik Sebzeler - Bakliyat - Ezme

Sepetler: Et & Et Ürünleri - Özel Sepetler - Ramazan Sepeti - Sepetler - Yılbaşı Sepeti - Reçel - Sebzeler - Granola - Salça

Temizlik ve Kişisel Bakım: Kişisel Bakım - Temizlik
Zeytin & Zeytinyağı: Zeytinyağı - Zeytin - Ezme

Kategoriye Göre Genel Satış Dağılımı



Şekil 8. Tüm satışların ürün kategorilerine göre dağılımı

Kategorilerle aynı isme sahip alt kategoriler, kategoriler arasında ortak alt kategoriler, ayrıca aynı kategorinin organik ve organik olmayan versiyonları olduğu görülmektedir. Eldeki satış verisinde mevcut olan ürünler üzerinden ürün verisi birleştirildiğinde 8 farklı kategori ve 27 farklı alt kategori görülmektedir. Bunlar, aşağıdaki gibidir:

Ekmek ve Unlu Mamüller: Ekmek - Yufka - Çikolata & Tatlı

Et, Tavuk, Balık: Et & Et Ürünleri - Tavuk - Balık & Deniz Ürünleri

Kahvaltılık: Peynir - Yumurta - Yoğurt - Süt Ürünleri - Et & Et Ürünleri - Sebzeler - Reçel - Kahvaltılık - Granola - Ekmek - Çikolata & Tatlı

Kuru Gıdalar: Kuruyemiş - Makarna, Manti & Un - Kuru Gıdalar - Ev Yemekleri - Çikolata & Tatlı - İçecekler

Meyveler: Meyveler

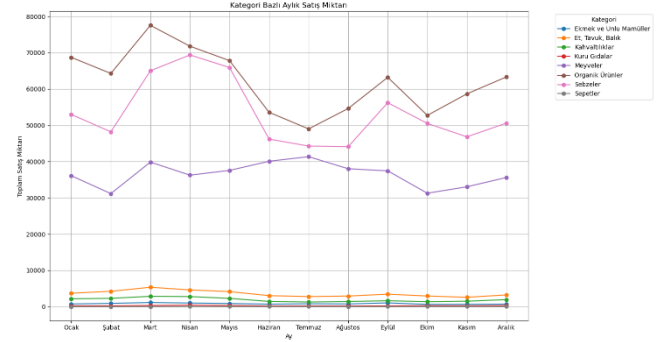
Organik Ürünler: Organik Sebzeler - Organik Meyveler - Organik Et, Tavuk - Organik Kahvaltılık - Organik Bal, Pekmez & Tahin

Sebzeler: Sebzeler - Bakliyat

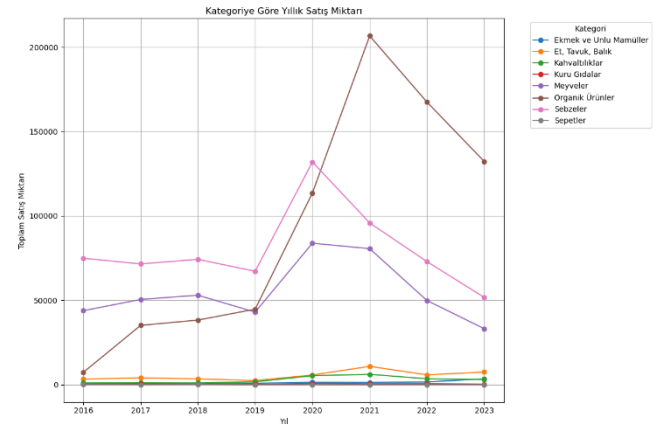
Sepetler: Özel Sepetler

Ürünlerin kategori bazlı aylık satışları incelenmiş, satışların meyve, sebze ve organik ürünlerde yoğunlaştığı ve satışların aylar üzerinden mevsimselliği Şekil 9'da sunulmuştur.

Ürünlerin kategori bazlı yıllık satış miktarları ise Şekil 10'da verilmiştir.



Şekil 9. Ürünlerin kategori bazlı aylık satışları



Şekil 10. Ürünlerin kategori bazlı yıllık satışları

Anomali İncelemesi

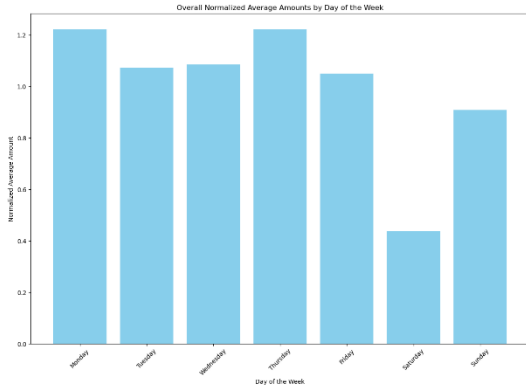
Satışlarda anomali incelemesi, tüm ürünlerin satış verisi üzerinden İzolasyon Ormanı tekniği kullanılarak yapılmıştır. Bu analiz sonucunda anomali olarak işaretlenen günlük satışlara rastlansa da, bunlar incelendiğinde sezon içerisinde ve genelde pandemi zamanında gerçekleştikleri görülmüştür. Kayıtlar incelendiğinde, bunların hatalı girdiler olduklarından şüphelenilmemiştir.

Bu inceleme sırasında ürün satışlarının çoğunlukla mevsimsel ve miktarlarının değişken olduğu, satışlar arasında mevsimsellikten ötürü veya ürünün satılmadığı yıl veya yılların olabildiği, bazı ürünlerin satışının sonlandığı ve bazı yeni ürünlerin de satılmaya başlandığı görülmüştür.

Eldeki veri kümesinin kapsadığı 2016-2023 yıllarına ait, talebe etkisi olabileceği düşünülen önemli tatil günleri, eklenen/bağlanan tatil günleri ve hafta sonları da dahil olmak üzere özel olarak tanımlanabilecek günler belirlendi. Satış verisi üzerindeki anomali incelemesine eklenerek anomali oluşturan günlerin bu günlerle kesişip kesişmediği incelendi. Ürünler mevsimsellikleri içinde gözlemlenirken, anomaliye dahil olan bu günlerin kendilerine özgü sebeplerle satışlara etki ettiğine dair bir bulguya rastlanmadı.

Haftanın günlerinin satışlar üzerindeki etkisi üzerine ayrı bir çalışma yapılmış ve her ürün için ayrı ayrı ve tüm ürünler birlikte olmak üzere ürünlerin satışlarının ne kadarının ne zaman gerçekleştiği incelenmiştir. Satış verisindeki tüm ürünler için yapılan bu çalışmada incelenen grafikler çok

fazla yer tuttuğu için makalede sunulmamıştır. Ürünlerin önemli bir kısmında cumartesi günlerinde satışın diğer günlere oranla gözle görülür şekilde düşük olduğu gözlemlenmiştir; aynı durum satışların normalize edildiği tüm ürünlere ait grafikte de gözlemlenmiştir (Şekil 11).



Şekil 11. Haftanın günleri bazında ortalama ürün satış miktarları (normalize edilmiş)

Verinin temizlenmesi

Eldeki satış verisinin eksik değerler barındırdığı, ancak temiz olduğu belirlenmiştir. Bunun üzerine, satış değerleri günlük olarak bir araya toplanmıştır. Satışın olmadığı günlere ilişkin kayıt bulunmadığından, bu günler için ürünler *Gerçek Ağırlık* = 0 olarak eklenmiştir. *Ürün Fiyatı* bilgisi yalnızca satış olduğu zamanlarda kaydedildiği ve ayrıca günlük değerlerinin kaydı bulunmadığı için, günlük gruplama sonrası oluşan eksik *Ürün Fiyatı* değerleri doldurulamamıştır. Bu durum, özellikle sık satılmayan ürünlerde yakın zamanlarda yaşanan fiyat değişikliklerinin ve etkilerinin takibini engellemiştir.

Eldeki kategori veri kümesinde eksik ve yanlış kayıtlar bulunduğu, ayrıca kategori ve alt kategori bilgilerinin doğrudan tutulmayıp dolaylı yoldan çıkarılması gerektiği belirlenmiş, gerekli çalışmalar yapılarak temiz ve eksiksiz bir veri kümesi elde edilmiştir.

Verinin Filtrelenmesi

Filtre: [*Toplama Türü* > 0] VE [*Gerçek Ağırlık* != 0] VE ~ [(*Gerçek Ağırlık* = 1) VE (*Tahmini Ağırlık* = (NaN))] (1)

Bu filtre (1), taze gıda olmayan ürünleri incelemekten hariç tutmak için uygulanmıştır. Filtreleme sonucunda 768 farklı kimlikli ürün bulunmuştur. Satışlar günler üzerinden birleştirilmiş, bir ürünün satışının olmadığı günlere ilişkin kayıt bulunmadığından, ürünün satışının olmadığı günlere ilişkin ilgili her ürün için *Gerçek Ağırlık* = 0 olarak eklenmiştir. Günlük kayıtların sayısı bu şekilde 2.170.368 (2826 gün * 768 ürün çeşidi) olmuş; bu da veriyi %86,53 oranında seyrek hale getirmiştir.

Verinin Ön İşlemesi

Yeni ürün tanıtımlarının işlenmesi: İncelenen zaman aralığında yeni ürün tanıtımları olduğu gözlemlendi. Bu ürünlerin dahil edilmesi çalışmada bilinmeyen sonuçlar doğurabileceği için, sonuçları etkilememek adına çıkarıldılar.

Kesintiye uğramış tarihsel satışların işlenmesi: Bazı ürünlerin uzun süre boyunca satışı olmadığı gözlemlendi. Bu

davranışlarının nedeni hakkında bilgi bulunmadığından, yanlışlıkları (*bias*) önlemek için, bu ürünler çalışmadan çıkarıldı.

Validasyon ve test kümelerini etkileyecek kesintiler:

Validasyon ve test kümelerini etkileyecek satış kesintilerine sahip ürünler de çalışmadan çıkarıldı. Bu ürünler, incelenen son iki yıldan en az birinde satışı olmayan ürünler olarak tanımlanmıştır.

Bu işlemler sonunda 159 farklı *Gıda Kimliği*'ne sahip ürün elde edildi. Bu verinin seyrekliği %53,26 olarak tespit edildi. Bu ürünlerin 7 farklı kategori ve 8 farklı alt kategoriye (- ile birbirlerinden ayrılmışlardır) ait olduğu görülmektedir:

Ekmek ve Unlu Mamüller: Ekmek

Et, Tavuk, Balık: Et & Et Ürünleri

Kahvaltılık: Peynir

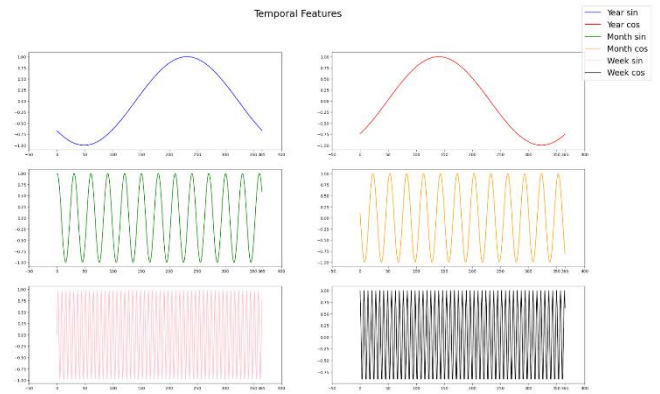
Kuru Gıdalar: Makarna, Mantı & Un

Meyveler: Meyveler

Organik Ürünler: Organik Sebzeler - Organik Meyveler

Sebzeler: Sebzeler

Tarihler, zaman damgalarına çevrildi ve Şekil 12'deki gibi haftalık, aylık ve yıllık periyotlu, artık yılları da dikkate alan sinüs ve kosinüs sinyallerine dönüştürülerek bir zamansal kodlama oluşturuldu.

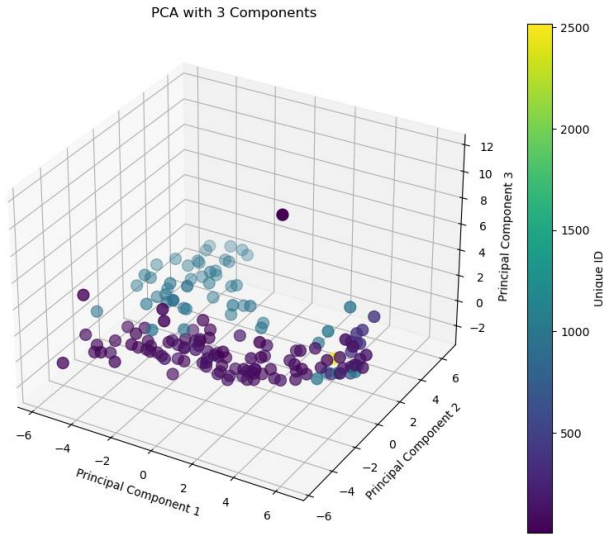


Şekil 12. Tarihlerle ait haftalık, aylık ve yıllık sinyaller

R programlama dilindeki *tsfeatures* paketinin Python dilindeki uygulanması kullanılarak eldeki zaman serilerinden özellikler çıkarılmış, ardından Ana Bileşen Analizi (*Principal Component Analysis*) 3 ana bileşen ile uygulanarak sonuçlar görselleştirilmiştir. Ayrı ayrı hem tüm zaman serilerine hem de ön işlemeden sonra seçilen zaman serilerine uygulanan bu yöntemde kullanılan özellikler şunlardır:

unitroot_pp, *unitroot_kpss*, *stability*, *trend*, *spike*, *linearity*, *curvature*, *e_acf1*, *e_acf10*, *x_pacf5*, *diff1x_pacf5*, *diff2x_pacf5*, *nonlinearity*, *lumpiness*, *alpha*, *beta*, *flat spots*, *entropy*, *crossing points*, *arch_lm*, *x_acf1*, *x_acf10*, *diff1_acf1*, *diff1_acf10*, *diff2_acf1*, *diff2_acf10*.

Şekil 13, seçilmiş zaman serilerinde yapılan çalışmayı görselleştirmektedir.

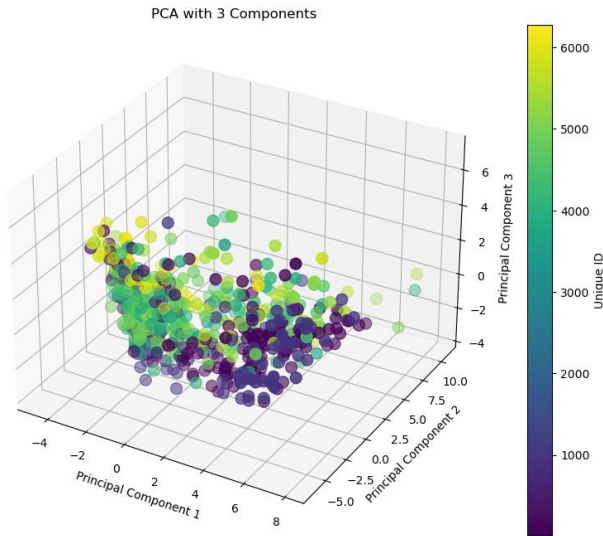


Şekil 13. Seçilmiş zaman serileri üzerinde 3 boyut ile PCA uygulanması sonucu ürünlerin uzayda görüntüsü

Ana bileşenler için açıkladıkları varyans oranları:

- Ana bileşen 1: 0.4558
- Ana bileşen 2: 0.1840
- Ana bileşen 3: 0.1080

ve toplam açıklanan varyans oranı 0.7478 olmuştur.



Şekil 14. Tüm zaman serileri üzerinde 3 boyut ile PCA uygulanması sonucu ürünlerin uzayda görüntüsü

Şekil 14, tüm zaman serilerinde yapılan çalışmaları görselleştirmiştir.

Ana bileşenler için açıkladıkları varyans oranları:

- Ana bileşen 1: 0.4255
- Ana bileşen 2: 0.1323
- Ana bileşen 3: 0.1043

ve toplam açıklanan varyans oranı 0.6621 olmuştur.

Yapılan bu analiz sonunda, zaman serilerinin incelenen özellikler açısından kolayca ayrılmadığı, ön işlemedeki zaman serisi seçimlerinin ise birbiriyle daha ilgili serilere ulaşmaya katkı sağladığı görülmektedir. Bunun üzerine TSNE ve UMAP boyut düşürme teknikleri de uygulanmış, sonuçları kümeleme işlemlerinde kullanılmıştır.

Kümeleme Çalışmaları

Çalışmanın bu safhasında, daha iyi sonuçlar elde edilmesine katkı sağlayabilecek zaman serileri kümeleri aranması üzerinde durulmuştur. Birbirine benzer zaman serilerinin ayırt edilebilmesi durumunda, her küme için farklı bir model eğitilerek tüm zaman serileri için tek bir model veya her zaman serisi için ayrı bir model düzenlerinden daha başarılı sonuçlar elde edilebileceği varsayılmıştır.

Seçilmiş zaman serilerinin verileri bu noktada eğitim, doğrulama ve test kümelerine bölünmüş, kümeleme çalışmaları ve daha sonraki eğitim çalışmalarında yalnızca eğitim verisi kullanılmıştır. 2826 günden ilk 2226 gün eğitim için ayrılmış, sonraki 300 gün doğrulama ve son 300 gün ise test için kullanılmıştır.

Python programlama dilindeki *tslearn* paketinden, zaman serileri üzerinde kümeleme görevleri için kullanılan *TimeSeriesKMeans* kümeleme metodu ve hizalanmamış zaman serileri için başarılı sonuçlar veren Dinamik Zaman Bükümü (*Dynamic Time Warping*) algoritması gibi önemli yapılar, *statsmodels* paketinden, bir zaman serisini trend, mevsimsellik ve artık bileşenlerine ayırmaya yarayan mevsimsel ayrıştırma (*seasonal decomposition*) metodu, *sktime* paketinden elde edilen zaman serisi verisini frekans alanına dönüştüren Ayırık Fourier Dönüşümü (*Discrete Fourier Transform*) araçlarından faydalanılarak yapılan kümeleme uygulamaları şu şekildedir:

Ham tek değişkenli zaman serisi verisi kullanılarak

- *TimeSeriesKMeans* kümeleme metodu uygulanması
- Hiyerarşik (*Hierarchical*) kümeleme metodu uygulanması
- Spektral (*Spectral*) kümeleme metodu uygulanması

Mevsimsel ayrıştırma ile

- Mevsimsellik bileşeni kullanılarak
 - *TimeSeriesKMeans* kümeleme metodu
 - Hiyerarşik (*Hierarchical*) kümeleme metodu
 - Spektral (*Spectral*) kümeleme metodu
- Trend bileşeni kullanılarak
 - *TimeSeriesKMeans* kümeleme metodu
 - Hiyerarşik (*Hierarchical*) kümeleme metodu
 - Spektral (*Spectral*) kümeleme metodu

Fourier Dönüşümü sonucunda elde edilen özellik kullanılarak

- *TimeSeriesKMeans* kümeleme metodu
- Hiyerarşik (*Hierarchical*) kümeleme metodu
- Spektral (*Spectral*) kümeleme metodu

tsfeatures paketi kullanılarak çıkarılan özellikler kullanılarak

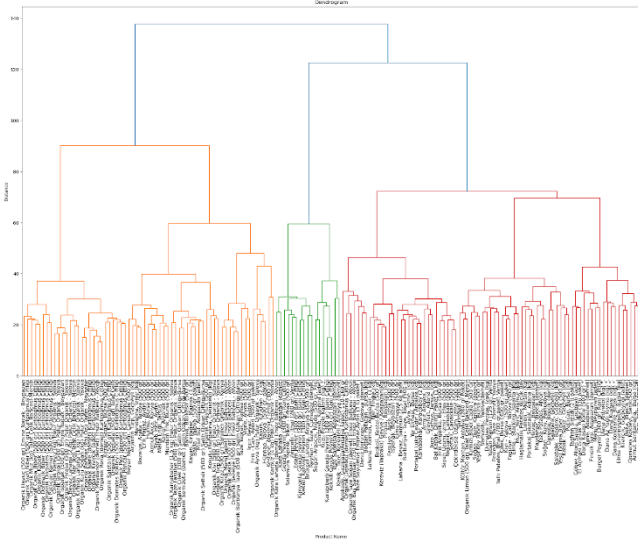
- *K-means* kümeleme metodu uygulanması
- Aglomeratif kümeleme metodu uygulanması
- Spektral (*Spectral*) kümeleme metodu uygulanması
- Gauss Karışım Modeli kümeleme metodu uygulanması
- DBSCAN kümeleme metodu kullanılması

Kümeleme başarıları silüet skoru (*silhouette score*) ile ölçülmüştür. Silüet skoru, bir kümeleme sonucunun

kalitesini değerlendirmek için kullanılan bir ölçüttür. Bir nesnenin kendi kümesiyle diğer kümelerle kıyasla ne kadar benzer olduğunu ölçer. Skor, -1 ile 1 arasında değişir: 1, nesnenin çok iyi kümelendiğini gösterir. 0, nesnenin iki komşu küme arasındaki karar sınırında veya ona çok yakın olduğunu gösterir. -1, nesnenin kötü kümelendiğini ve yanlış kümeye atanmış olabileceğini gösterir.

Tüm nesneler için ortalama silüet skoru hesaplanır. Daha yüksek bir ortalama skor genellikle daha iyi bir kümeleme kalitesi anlamına gelir. 0.5'in üzerindeki skorlar genellikle iyi olarak kabul edilirken, 0.7'nin üzerindeki skorlar çok iyi bir kümelemeyi gösterir.

Ham veri üzerinden yapılan tüm kümelemelerde, benzerlik ölçütü olarak Dinamik Zaman Bükümü (*Dynamic Time Warping*) algoritması kullanılmıştır. Satış dönemlerinde meydana gelebilecek potansiyel bir kaymanın sonuçlara etki etmemesi için noktalar arası Öklid mesafesi (*Euclidean distance*) gibi ölçütler yerine tercih edilmiştir. Mevsimsel ayrıştırma bir zaman serisini, verinin davranışını daha iyi anlamak için temel bileşenlerine ayırmak amacıyla kullanılan bir tekniktir. Bir zaman serisi üç ana bileşene ayrılır:



Şekil 15. Hiyerarşik kümeleme sonuçları

Trend: Bu bileşen, verideki uzun vadeli ilerlemeyi temsil eder. Verinin zamanla yukarı, aşağı veya yatay hareket edip etmediğini gösteren genel yönelimdir. Trend bileşeni, kısa vadeli dalgalanmaları hariç tutarak verinin altında yatan deseni belirlemeye yardımcı olur.

Mevsimsellik: Bu bileşen, verideki tekrarlayan ve periyodik karakterli desenleri yakalar. Bu desenler, mevsimsel faktörler nedeniyle belirli aralıklarla düzenli olarak ortaya çıkar. Örneğin, satışlar yaz tatilinde insanların evlerinde bulunmaması nedeniyle azalabilir veya belirli ürünlere olan talep belirli mevsimlerde yükselir.

Artık veya Düzensiz Bileşen: Bu bileşen, trend veya mevsimsellik tarafından açıklanamayan verideki gürültü veya rastgelelikleri içerir. Bu, öngörülebilir bir deseni takip etmeyen rastgele dalgalanmalar veya düzensiz değişikliklerdir.

Çalışma sonucu alınan kümeleme sonuçları aşağıda belirtilmiştir:

- Ham tek değişkenli zaman serisi verisi kullanılarak 0,5 üzerinde silüet skoruna sahip bir kümeleme bulunamadı.
- Mevsimsel ayrıştırma ile mevsimsellik bileşeni kullanılarak 0,5 üzerinde silüet skoruna sahip bir kümeleme bulunamadı. En yüksek ikinci silüet skorları bu bileşen ile elde edildi.
- Mevsimsel ayrıştırma ile trend bileşeni kullanılarak 0,5 üzerinde silüet skoruna sahip bir kümeleme bulunamadı.
- Fourier Dönüşümü sonucunda elde edilen özellik kullanılarak 0,5 üzerinde silüet skoruna sahip bir kümeleme bulunamadı.
- *tsfeatures* üzerinden çıkarılan özellikler kullanılarak 0.5 üzerinde silüet skoruna sahip kümelemeler bulundu.

Ham zaman serisi verisi üzerinden elde edilen Hiyerarşik Kümeleme sonuçları Şekil 15'te gösterilmiştir.

Spektral kümeleme sonuçlarının tamamının sonuçları 0'ın altında silüet skoru almış olup kötü kümelenebilir işaret etmektedir. Bu durum verinin tek değişkenliğinden veya karmaşık bir küme yapısının olmamasından kaynaklanabilir.

Mevsimsellik bileşeninin ham veriyi kümelemeye en elverişli bileşen olduğu tespit edilmiş, Dinamik Zaman Bükümü algoritmasının geliştirilmiş bir versiyonu olan *SoftDTW* ile kümeleme işlemi en yüksek silüet skorunu alan *TimeSeriesKMeans* metodu ile tekrar çalıştırıldı. 0.5 üzerinde silüet skoruna sahip kümelemeler bulundu. *SoftDTW* metodunun hesaplama karmaşıklığının yüksek olduğu, özellikle de kullanılan veri kümesi gibi uzun süreli zaman serilerinde çok uzun sürede sonuç verdiği not edilmektedir.

Dinamik Zaman Bükümü (*Dynamic Time Wrapping-DTW*) ve raporda açıklandığı üzere ilgili alanlarda geliştirilmiş versiyonu olan *SoftDTW* algoritmalarından alınan kümeleme sonuçları aşağıda verilmiştir.

Ham veri üzerinde *TimeSeriesKMeans* kümelemesinin, her Gıda Kimliği için tek seri kullanılarak, tüm süre üzerinden normalize edilmiş satış miktarları üzerinden benzerlik sonuçları Tablo 1'de, hiyerarşik ve spektral kümeleme sonuçları ise Tablo 2'de verilmiştir.

Tablo 1. Ham veri üzerinde TimeSeriesKMeans kümelemesi sonuçları

DTW		SoftDTW	
Küme Sayısı	Silüet Skoru	Küme Sayısı	Silüet Skoru
2	0.166767	2	0.236162
3	0.129943	3	0.173653
4	0.080402	4	0.118087
5	0.075215	5	0.122781

6	0.082058	6	0.153005	Küme Sayısı	Silüet Skoru	Küme Sayısı	Silüet Skoru
Tablo 2. Ham veri üzerinde hiyerarşik ve spektral kümeleme sonuçları				2	0.394005	2	-0.015400
				3	0.325798	3	-0.035023
				4	0.324963	4	-0.063858
				5	0.316500	5	-0.076701
				6	0.329073	6	-0.101301

Tablo 5. Trend bileşeni üzerinde *TimeSeriesKMeans* kümelemesi sonuçları

DTW	
Küme Sayısı	Silüet Skoru
2	0.202558
3	0.173458
4	0.172085
5	0.177106
6	0.165388

Tablo 6. Trend bileşeni üzerinde hiyerarşik ve spektral kümeleme sonuçları

Hiyerarşik kümeleme (DTW)		Spektral kümeleme (DTW)	
Küme Sayısı	Silüet Skoru	Küme Sayısı	Silüet Skoru
2	0.208320	2	-0.018545
3	0.168085	3	-0.038094
4	0.180301	4	-0.055345
5	0.168528	5	-0.068271
6	0.167159	6	-0.084200

Tablo 7. Fourier dönüşümü çıktısı üzerinde *TimeSeriesKMeans* kümelemesi sonuçları

DTW	
Küme Sayısı	Silüet Skoru
2	0.228800
3	0.228623
4	0.145337
5	0.175208

Tüm süre üzerinden normalize edilmiş satış miktarlarına sahip serilerden mevsimsel ayrıştırma ile elde edilen mevsimsellik ve trend bilgileri üzerinden kümeleme yapılmış, her ürün için tek seri kullanılmıştır. Mevsimsellik için Tablo 3'te *TimeSeriesKMeans* kümeleme algoritması ile uzaklık hesaplamaları için DTW ve *SoftDTW* kullanılması sonucu alınan silüet skorları, Tablo 4'te hiyerarşik ve spektral kümeleme algoritmaları ile DTW kullanılması sonucu alınan silüet skorları verilmiştir. Trend için Tablo 5'te *TimeSeriesKMeans* kümeleme algoritması ile DTW kullanılması sonucu alınan silüet skorları, Tablo 6'da hiyerarşik ve spektral kümeleme algoritmaları ile DTW kullanılması sonucu alınan silüet skorları verilmiştir. Ayrıca her ürün için tek seri kullanılarak Fourier dönüşümü de uygulanmış, çıktılar üzerinde DTW kullanılmış, çıktılar Tablo 7 ve Tablo 8'de verilmiştir.

Tablo 3. Mevsimsel bileşen üzerinde *TimeSeriesKMeans* kümelemesi sonuçları

DTW		SoftDTW	
Küme Sayısı	Silüet Skoru	Küme Sayısı	Silüet Skoru
2	<u>0.398851</u>	2	<u>0.562553</u>
3	0.409759	3	0.574747
4	0.342410	4	0.513856
5	0.331625	5	0.512557
6	0.320067	6	0.508367

Tablo 4. Mevsimsel bileşen üzerinde hiyerarşik ve spektral kümeleme sonuçları

Hiyerarşik kümeleme (DTW)		Spektral kümeleme (DTW)	
---------------------------	--	-------------------------	--

6

0.162159

DBSCAN

5²

0.472211

Tablo 8. Fourier dönüşümü çıktısı üzerinde hiyerarşik ve spektral kümeleme sonuçları

Hiyerarşik kümeleme (DTW)		Spektral kümeleme (DTW)	
Küme Sayısı	Silüet Skoru	Küme Sayısı	Silüet Skoru
2	0.173000	2	-0.008773
3	0.207786	3	-0.025966
4	0.151363	4	-0.052478
5	0.152575	5	-0.130124
6	0.125460	6	-0.136958

İncelenen tüm *tsfeatures* özellikleri üzerinden veya PCA ve TSNE boyut düşürme yöntemleri ile 0,5 üzerinde silüet skoruna sahip kümeleme bulunamasa da, UMAP boyut düşürme metodu ile 0.5 üzerinde silüet skoruna sahip kümelemeler bulundu. Bu aday kümelemeler üzerinden Davies-Bouldin Index ve Calinski-Harabasz Index değerleri incelenerek en iyi adaylar seçildi. Sonuçlar Tablo 9'da gösterilmiştir.

LSTM modeli, bu çalışmadaki veri üzerinden hiperparametre ayarlamasından geçirilmiş, iki katmanlı, katman çıktısından sonra Dropout uygulanan, her katmandaki hücre sayısı 448 olan bir ağı en yüksek başarıyı gösterdiği belirlenmiştir. Aynı şekilde öğrenme hızı parametresi de seçilmiştir. Çalışmalarda tek *Gıda Kimliği* üzerinden eğitilmeyen modellerde, LSTM durumlu (*stateful*) olarak kullanılmış, parti boyutu buna göre ayarlanmış, partiler uygun şekilde verilmiş ve karıştırma engellenmiştir. Tek *Gıda Kimliği* üzerinden eğitilen modellerde parti boyutu 32 olarak seçilmiştir.

Tablo 9. *tsfeatures* özelliklerinden UMAP boyut düşürme yöntemi ile alınan en yüksek silüet skorlu kümeleme sonuçları

Kümeleme Yöntemi	Küme Sayısı	Silüet Skoru
K-means	3 ¹	0.699991
K-means	4	0.663826
Aglomeratif	3	0.699991
Aglomeratif	4	0.647261
Spektral	3	0.699991
Spektral	4	0.661110
GMM	3	0.699991
GMM	4	0.651314

¹ 3 kümeli yaklaşımların tamamında aynı kümelemenin yapıldığı görülmüştür.

² DBSCAN tüm veriyi kümeleme zorunluluğu olmadığı için 5 küme bulmuş olup bu kümelere yerleştirmede birçok aykırı veri de bulunmaktadır.

Bu model ile önce tüm zaman serisi verisinin ham haliyle tek model ile eğitildiği ve her *Gıda Kimliği* zaman serisi verisi için ayrı bir modelin eğitildiği iki farklı çalışma yapılmıştır. Tek model ile alınan sonuçların daha başarılı olduğu görülmüştür. Bu durum, literatürdeki bulgularla da örtüşmektedir. Bu çalışmadan referansla birbirine benzer zaman serilerinin bir arada kümelmesiyle daha başarılı sonuçlar elde edilip edilemeyeceği incelenmek istenmiştir. Yapılan kümeleme çalışmalarında kümelemeye uygun sonuçlar seçilerek her küme için ayrı bir model eğitilmiştir. Alınan sonuçlarda istatistiksel olarak önemli miktarda iyileşme sağlayan bir kümeleme bulunamamıştır. Ancak veride belirtilen kategori ve alt kategorilere göre yapılan bir kümeleme ile karşılaştırıldığında daha iyi sonuç veren kümelemelere rastlanmıştır.

XGBoost modeli, bu çalışmada erişilen kategori bilgileri de dahil edilerek kullanılmıştır. XGBoost'un kategorik verileri işleme yeteneğinin olması ve bu yeteneğin kolayca etkinleştirilebilmesi, kategori çalışmasının bu modelde yapılmasını istenir kılmalıdır. Üç ayrı çalışma yapılmış, ilkinde *Gıda Kimliği* bilgileri, ikincisinde hem *Gıda Kimliği* hem de Kategori ve Alt Kategori bilgileri, XGBoost modeline ham verinin yanında kategorik özellik olarak verilmiş, üçüncü ise yalnızca ham veri olarak eğitilmiştir. Yalnızca *Gıda Kimliği*'nin eklenmesi sonucunda istatistik olarak önemli boyutta bir iyileşme gözlenmiş; ancak Kategori ve Alt Kategori bilgilerinin etkilenmesi performansı daha fazla etkilememiştir. Bu durum eldeki Kategori ve Alt Kategori bilgilerinin oldukları gibi kullanılmayıp öğrenmeye fayda sağlaması için ayrıca işleme gerektirebileceğini düşündürmektedir.

Eğitim Sonuçları

Tüm eğitimler 5 numune ile gerçekleştirilmiş, ortalama ve standart sapma değerleriyle birlikte (XGBoost modelinde stokastik parametre seçimi bulunmasına rağmen verinin aynı sırada verilmesi aynı sonuçların alınmasını sağlamıştır), MSE ve MAE değerleri Tablo 10'da sunulmuştur.

Tablo 10. Ham veri üzerinde kümelemesiz (tek küme) düzende LSTM ve XGBoost model sonuçları

	LSTM	XGBoost	XGBoost (+FoodId) ³
MSE	13.437 ± 0.154	14.085	13.629
MAE	1.786 ± 0.008	1.830	1.792

³ LSTM modeli *stateful* olarak kullanıldığı için örtük olarak hangi verinin hangi seriden geldiği bilgisine sahiptir. XGBoost modelinin de bu avantajdan faydalanabilmesi için *FoodId* bilgisi kategorik olarak eklenmiş, performansı LSTM seviyesine getirdiği görülmüş, raporlanan sonraki modellerde bu şekilde kullanılmıştır.

159 küme düzeni üzerinde LSTM ve XGBoost modellerinden alınan sonuçlar Tablo 11’de sunulmuştur.

Tablo 11. Ham veri üzerinde her Gıda Kimliği üzerinde ayrı kümeleme düzeninde LSTM ve XGBoost model sonuçları

	LSTM	XGBoost
MSE	15.262 ± 0.257	17.032
MAE	1.933 ± 0.011	2.040

En yüksek silüet skoru alan mevsimsel bileşen üzerinde yapılan kümelemelerle uygulanan eğitimin sonuçları Tablo 12’de verilmiştir.

Tablo 12. Mevsimsel bileşen üzerinde *TimeSeriesKMeans* kümeleme düzeninde LSTM ve XGBoost model sonuçları

	LSTM (2 küme)	LSTM (3 küme)	XGBoost (2 küme)	XGBoost (3 küme)
MSE	13.663 ± 0.366	13.195 ± 0.179	13.446	13.378
MAE	1.821 ± 0.009	1.807 ± 0.022	1.798	1.783

En yüksek silüet skoru alan *tsfeatures* özellikleri üzerinde UMAP boyut düşürme metodu uygulandıktan sonra yapılan kümelemelerle uygulanan eğitimin sonuçları Tablo 13’te verilmiştir. Verideki kategori ve alt kategori bilgisine göre yapılan kümeleme ve kümeleme uygulanmama durumu ile aynı seviyede sonuçlar alınmıştır.

Tablo 13. Uygulanan kümeleme yöntemlerinin en iyileri kullanılarak alınan tahmin sonuçları

Model	Boyut Düşürme Yöntemi	Kümeleme Yöntemi (Küme Sayısı)	MSE	MAE
LSTM	<i>tsfeatures</i> özellikleri üzerinden UMAP	DBSCAN (5)	14.344 ± 1.4955	1.930 ± 0.132
XGBoost			13.538	1.809
LSTM		K-means (3) ⁴	13.335 ± 0.392	1.793 ± 0.023
XGBoost			13.407	1.786
LSTM		K-means (4)	12.995 ± 0.131	1.796 ± 0.009
XGBoost			13.411	1.789
LSTM		Aglomeratif kümeleme (4)	15.170 ± 0.821	1.933 ± 0.054
XGBoost			13.453	1.791

⁴ Aynı kümeleme çıktısına aglomeratif kümeleme (3), spektral kümeleme (3) ve GMM kümelemesi (3) yöntemleri de sahiptir.

LSTM		Spektral kümeleme (4)	13.004 ± 0.162	1.805 ± 0.024
XGBoost			13.435	1.789
LSTM		GMM kümeleme (4)	13.039 ± 0.116	1.802 ± 0.007
XGBoost			13.401	1.788
LSTM	Yok	Kategori ve alt kategori bilgisine göre (8)	14.755 ± 2.079	1.886 ± 0.138
XGBoost			13.393	1.795
LSTM		Kategori ve alt kategori bilgisinin girdiye eklenmesiyle (kümeleme yok)	13.422 ± 0.547	1.785 ± 0.016
XGBoost			13.601	1.79

Kapsamlı kümeleme çalışmalarından fayda sağlanamaması sonrasında, tüm ürünlerin verisinin tek bir modelde kullanıldığı yoldan devam edilmesi tercih edilmiş, ürünlerdeki indirim bilgisi girdilere eklenmiştir. Tahmin edilecek günün satış değerinin indirimden etkileneceği için o günün zamansal özelliklerinin de indirim bilgisiyle beraber modele girdi olarak verilmesi uygun bulunmuştur. 7 günlük geçmiş veri girdisiyle yapılan sonraki gün tahminleri; tahmin yapılacak gün için indirim bilgisinin eklenmesinin gerekliliği üzerine indirim bilgisinin ve tahmini yapılacak günün zamansal özelliklerinin de eklenmesi, girdiler arası boyut uyumu için de tahmin gününde gerçek talep miktarının yerine farklı bir değer kullanılması şeklinde yapılmıştır. Modeller yeni girdiler ile baştan eğitilmiş ve alınan sonuçlar önceki sonuçlarla karşılaştırılmıştır. Bu eğitimin sonuçları Tablo 14’te verilmiştir.

Tablo 14. Kümeleme kullanılmayan yeni girdi modeli ile karşılaştırma ve yeni girdi modeli + indirim sonuçları

Model	Girdi Modeli	MSE	MAE
LSTM	7 + 1 (ortalama değer)	14.180 ± 0.724	1.808 ± 0.027
XGBoost		13.751	1.8
LSTM	7 + 1 (son değer)	13.180 ± 0.074	1.761 ± 0.020
XGBoost		13.767	1.809
LSTM	7 + 1 (son değer) + indirim bilgisi	12.209 ± 0.270	1.692 ± 0.030
XGBoost		12.541	1.738

Yeni girdi ile eğitilen modellerin kendi başlarına istatistiksel olarak önemli bir avantajları olmadığı görülmüş, indirim bilgisinin eklenmesinin ise hatada istatistiksel olarak önemli bir azalma sağladığı tespit edilmiştir.

Sonuç

Bu çalışma, e-ticaret sektöründe faaliyet gösteren gıda perakendecileri için daha doğru talep tahminleri yapmayı amaçlayan bir sistem geliştirmiştir. Günümüzde işletmeler için stok yönetimi ve tedarik zinciri optimizasyonu gibi süreçler kritik öneme sahiptir. Özellikle gıda sektöründe, hızla bozulabilen ürünlerin talebinin doğru tahmin edilmesi, işletmelerin bu süreçlerde maliyetlerini düşürmesine ve müşteri memnuniyeti ve sürdürülebilirliği artırmasına yardımcı olmaktadır. Çalışmamızda, 2016-2023 yılları arasındaki yedi yıllık satış verileri analiz edilerek taleplerin nasıl değiştiği ve yakın gelecekte nasıl öngörülebileceği incelenmiştir. Çalışmanın katkıları şu şekilde özetlenebilir: Makine öğrenmesi modelleri (LSTM ve XGBoost) ile satış tahmini yapılması, veriyi daha anlamlı hale getirmek için kümeleme yöntemlerinin uygulanması ve pandemi gibi olağanüstü durumların talep üzerindeki etkisinin analiz edilmesi.

Yapılan testler sonucunda, LSTM modeli en başarılı tahmini yapmış olduğu, XGBoost modelinin de aynı seviyede güçlü olduğu ve hızı sebebiyle etkili bir alternatif olduğu sonucu çıkarılmıştır. Kümeleme yöntemleri incelendiğinde, mevsimselliğe dayalı gruplandırmaların daha anlamlı tahminler üretebildiği gözlemlenmiştir. Ayrıca, ürünlerdeki fiyat indirimleri gibi dışsal faktörlerin tahmin modellerine dahil edilmesinin, tahmin doğruluğunu artırabileceği belirlenmiştir.

Bu çalışmanın ortaya koyduğu metodoloji, yalnızca gıda sektörüyle sınırlı kalmadan, e-ticaretin farklı alanlarında da uygulanabilir. Benzer bir yaklaşımla, elektronik, moda ve kozmetik gibi sezonsallık barındıran sektörlerde de talep tahminlerini iyileştirmek mümkündür. Zira, pandemi gibi sıradışı bir dönemde bazı sektörlerdeki müşteri hareketlerinde artış gözlenirken (örneğin online gıda), başka sektörlerdeki hareketlerde azalma (örneğin kozmetik) görülmüştür. Sonuç olarak, bu çalışma, işletmelerin daha iyi kararlar almasını sağlamak ve talep tahmin süreçlerini optimize etmek için veri odaklı genel bir çözüm sunmaktadır. Gelecekte, farklı sektörlerde benzer analizlerin yapılması ve daha fazla dış faktörün (özel günler, hava durumu, sosyal eğilimler vb.) modellenmesi, bu tür tahmin sistemlerini daha da geliştirilebilir.

Teşekkür

Bu araştırma, Tübitak 1501 programı destekli 3230999 numaralı 'FoReML: Gıda Perakende Sektöründe Talep Tahmin ve Öneri Sistemi' projesi ve Galatasaray Üniversitesi FBA-2024-1258 numaralı BAP projesi tarafından desteklenmiştir.

Yazar Katkıları

- Çalışma konsepti ve tasarım: Bozdoğan, Alptekin
- Veri toplama: Bozdoğan
- Verilerin analizi ve yorumlanması: Bozdoğan
- Taslağın oluşturulması: Bozdoğan, Alptekin
- Revizyon: Alptekin

Kaynaklar

- [1] K. Lutoslawski, M. Hernes, J. Radomska, M. Hajdas, E. Walaszczyk, and A. Kozina, "Food demand prediction using the nonlinear autoregressive exogenous neural network," *IEEE Access*, vol. 9, pp. 146 123–146 136, 2021.
 - [2] Analytics Vidhya. (2020). Food Demand Forecasting [Data set]. Kaggle.
<https://doi.org/10.34740/KAGGLE/DSV/1138962>
 - [3] M. Aci and D. Yergök, "Demand forecasting for food production using machine learning algorithms: A case study of university refectory," *Tehnikivjesnik*, vol. 30, no. 6, pp. 1683–1691, 2023.
 - [4] F. T. Liu, K. M. Ting, and Z.-H. Zhou, "Isolation forest," in *2008 Eighth IEEE International Conference on Data Mining*, 2008, pp. 413–422.
 - [5] D. Müllner, "Modern hierarchical, agglomerative clustering algorithms," *arXiv preprint arXiv:1109.2378*, 2011.
 - [6] A. Ng, M. Jordan, and Y. Weiss, "On spectral clustering: Analysis and an algorithm," in *Advances in Neural Information Processing Systems*, vol. 14, 2001.
- Tablo 13. Uygulanan kümeleme yöntemlerinin en iyileri kullanılarak alınan tahmin sonuçları