



Müşteri Segmantasyonu İle Karar Destek Sistemi

Onur Arslan , Kağan Deniz , Aybüke Kılıçsaymaz

Muğla Sıtkı Koçman Üniversitesi, Bilişim Sistemleri Mühendisliği Bölümü, Muğla

Özet

Bu proje, müşterilerin demografik ve ekonomik verilerini kullanarak makine öğrenimi tabanlı bir segmentasyon yaklaşımı sunmaktadır. Random Forest sınıflandırma modeli ile müşteriler belirli öbeklere ayrılır, bu segmentlere özel reklam stratejisi ve alışveriş kredisi teklifleri oluşturulur. Kullanıcıdan elde edilen yeni veriler modele eklenerek zamanla verisetinin genişlemesi ve modelin yeniden eğitilmesi sağlanır. Böylece, uygulama sürekli öğrenerek pazarlama stratejilerinin daha kişiselleştirilmiş ve etkili hale gelmesine katkıda bulunur.

Anahtar Kelimeler: *Müşteri Segmantasyonu , Makine Öğrenmesi , Veri Analizi*

Decision Support System with Customer Segmentation

Abstract

This project employs a machine learning-based segmentation approach to categorize customers into distinct clusters using demographic and economic data. Leveraging a Random Forest classification model, customers are grouped into predefined segments, and personalized advertising strategies and shopping credit offers are generated for each cluster. As new user data is continuously integrated, the dataset expands, enabling the model to be retrained over time. Consequently, the application continually improves, resulting in more personalized and effective marketing strategies.

Keywords: *Customer Segmantation, Machine Learning, Data Analysis,*

1 Giriş

Günümüzde işletmeler, müşteri segmentasyonunu etkin bir biçimde kullanarak pazarlama stratejilerini daha verimli kılmayı ve müşteri memnuniyetini artırmayı amaçlamaktadır. Özellikle dijital platformların yaygınlaşması ve veri miktarının hızla artması, tüketici davranışlarının daha ayrıntılı analiz edilmesine ve kişiselleştirilmiş pazarlama hamlelerinin hayata geçirilmesine olanak tanımaktadır. Bu doğrultuda, müşterilerin demografik özellikleri, gelir düzeyleri, istihdam durumları, satın alma alışkanlıkları ve ilgi alanları gibi bir dizi değişkenin kullanıldığı veri odaklı segmentasyon çalışmaları önem kazanmaktadır.

Bu proje kapsamında geliştirilen uygulama, bir makine öğrenimi modeli (Random Forest) kullanarak müşterileri belirli öbeklere 4(segmentlere) ayırmayı hedeflemektedir. Bu sayede, kullanıcıdan alınan basit giriş verileriyle müşterinin hangi öbeğe ait olabileceği tahmin edilerek, ilgili müşteri tipine uygun reklam stratejileri, ürün önerileri ve alışveriş kredisi teklifleri sunulabilmektedir. Proje ayrıca, yeni kullanıcı tahminlerini kaydederek veri setinin genişletilmesine ve modelin zaman içinde yeniden eğitilerek güncel kalmasına da imkân sağlamaktadır.

Bu raporun ilerleyen bölümlerinde, geliştirilen uygulamanın teknik altyapısı, veri ön işleme adımları, model oluşturma süreci, ürün ve strateji öneri mekanizması, yeni verilerin eklenmesi ve modelin yeniden eğitimi süreçleri detaylı biçimde ele alınacaktır. Böylelikle, uygulamanın hem teorik temelleri hem de pratik kullanım alanları bütüncül bir bakış açısıyla aktarılacaktır.

2 Literatür Taraması

Budak, H., & Gümüştaş, E. [1] Tarafından yapılan çalışmada, müşterilere kişiselleştirilmiş ürün önerileri sunmayı amaçlayan bir sistem geliştirilmiştir. Firmalar, müşterilerin tercihlerine göre öneriler sunarak hem satışlarını artırmak hem de müşteri memnuniyetini yükseltmek istiyor. Makale, işbirlikçi filtreleme yöntemini geliştirmek için k-means kümeleme algoritmasını işbirlikçi filtreleme ile birleştirerek hibrit bir yöntem

önermiş. K-means sayesinde, benzer özelliklere sahip kullanıcılar alt gruplara ayrılmış ve her gruba özel öneri modelleri oluşturulmuş. Sonuç olarak, önerilen yöntemin hata oranı daha düşük bulunmuş, yani önerilerin isabet oranı artmış.

Sinap, V. (2024) [2]. Tarafından yapılan çalışmada, perakende sektöründe Black Friday satışlarını tahmin etmek amacıyla kullanılan farklı makine öğrenmesi algoritmalarını kapsamlı bir şekilde karşılaştırıyor. Çalışmada, Doğrusal Regresyon (LR), Rastgele Orman (RF), K-En Yakın Komşu (KNN), XGBoost (XGB), Karar Ağacı (DT) ve LightGBM (LGBM) gibi yaygın kullanılan makine öğrenmesi algoritmaları kullanılarak satış tahminlemesi yapılmıştır. Araştırmanın temel amacı, bu algoritmaların performanslarını farklı değerlendirme metrikleri kullanarak karşılaştırmak ve en iyi tahminleme modelini belirlemektir.

İlk olarak, veriler üzerinde Keşifsel Veri Analizi (EDA) gerçekleştirilmiş ve bu süreçte müşteri davranışları, demografik özellikler ve satın alma eğilimleri detaylı bir şekilde incelenmiştir. EDA sürecinde elde edilen bulgular, verilerin makine öğrenmesi modellerine en uygun şekilde entegre edilmesine katkı sağlamış ve modelleme aşamasının daha verimli yapılmasına olanak tanımıştır. Özellikle müşteri yaş grupları, gelir düzeyleri ve satın alma sıklığı gibi özellikler üzerinden yapılan analizler, modellerin daha isabetli tahminlerde bulunmasını sağlamıştır.

Ergun, O. [3] tarafından yapılan çalışmada, e-ticaret verileri üzerinden müşterilere daha etkili hizmet sunmayı amaçlayan bir müşteri segmentasyonu sistemi geliştirilmiştir. Bu çalışmada, müşterilerin alışveriş alışkanlıklarına dayalı olarak gruplandırılması ve her gruba özel stratejiler geliştirilmesi hedeflenmiştir. Bu kapsamda, Random Forest algoritması ve k-means kümeleme yöntemi bir arada kullanılarak bir hibrit yaklaşım önerilmiştir.

Random Forest algoritması, veri setindeki önemli değişkenlerin seçilmesi ve analiz edilmesi için kullanılmıştır. Bu süreç, müşteri davranışlarının daha iyi anlaşılmasını sağlamış ve analiz edilen değişkenlerle güçlü bir temel oluşturulmuştur. Sonrasında, k-means algoritması ile müşteriler, benzer alışkanlıklara sahip oldukları yedi farklı gruba ayrılmıştır. Her grup, belirli alışveriş

özelliklerine göre karakterize edilmiş ve bu gruplara özel stratejiler geliştirilmiştir.

Sonuçlar, Random Forest ve k-means algoritmalarını birleştiren bu hibrit yöntemin, müşteri segmentasyonu süreçlerinde doğruluğu artırdığını göstermektedir. Özellikle, her bir grup için özel önerilerin geliştirilmesi, müşteri memnuniyetinin artırılması ve işletme karlılığının yükseltilmesi açısından önemli katkılar sunmaktadır.

3 Materyal ve Yöntem

Bu çalışmada 9 adet kategorik 4 adet sürekli değişkenin sınıflandırma başarısını ölçmek için Random Forest , SVM , Decision Tree , Gradient Boost , Logistic Regression , Gaussian Nb ve KNN yöntemleri test edilmiştir bu yöntemler aşağıdaki başlıklarda açıklanacaktır.

3.1 Makine Öğrenmesi

Makine öğrenmesi, bilgisayarın meydana gelen bir olay ile ilgili edindiği bilgileri ve tecrübeleri öğrenmek suretiyle, gelecekte oluşabilecek benzer olaylar hakkında kararlar verebilmesi ve oluşacak problemlere çözüm üretebilmesidir (Öztemel E., 2006)[4]. Makine öğrenmesi bazı yöntemler kullanarak geçmişteki verilerden yararlanır ve yeni veri için en uygun modeli bulmaya çalışır.

Çok büyük miktarlardaki verinin elle işlenmesi ve analizinin yapılması oldukça zordur. Burada amaç geçmişteki verileri kullanarak gelecek durumlar için tahminlerde bulunabilmektir. Uygulama alanı ne olursa olsun, çok miktardaki verinin analiz edilerek gelecek ile ilgili tahminlerde bulunması ve bizim karar vermemize yardımcı olması sayesinde makine öğrenmesi yöntemlerinin önemi her geçen gün artmaktadır.

Yapay Zekânın alanının bir dalı olan Makine Öğrenmesi (Machine Learning), bilgisayarların “öğrenme” görevini yerine getirecek algoritma ve tekniklerin gelişimi ile ilgilenir. Makine öğrenmesi; Doğal Dil işleme, Konuşma ve El Yazısı Tanıma, Nesne Tanıma, Bilgisayar Oyunları, Robot Hareketleri, Arama Motorları ve Tıbbi Teşhis gibi birçok alanda kullanılır (Kutlugün ve ark., 2017).

Makine öğrenmesinin üç önemli aşaması vardır. Bunlar:

1. Dokümanların Hazır Hale Getirilmesi

2. Öğrenme Yöntemlerinin Kararlaştırılması ve Uygulanması

3. Öğrenme Performansının Değerlendirilmesi

Makine öğrenmesinde öncelikle öğrenme işleminin yapılacağı veri kümesinin uygulanacak öğrenme yöntemine uygun halde hazırlanması gerekmektedir. Öğrenme metodunda istatistiksel yöntemler kullanılmaktadır. Yeni geliştirilen metotlar da

istatistiksel temellidir. Yeni bir metot bulunduğunda bu metodun performansı ölçülmektedir. Bu sayede diğer metotlarla karşılaştırılması yapılmaktadır Makine öğrenmesi metotlarını farklı uygulamaları analiz etmek ve farklı beklentilere göre sınıflandırmak mümkündür

3.2 Sınıflandırma yöntemleri

3.2.1 Random Forest

Random Forest, makine öğrenmesinde sıkça kullanılan güçlü ve esnek bir denetimli öğrenme algoritmasıdır. Temelde, birçok karar ağacının (decision tree) birleşiminden oluşan bir topluluk yöntemidir. Her ağaç, eğitim verisinin rastgele bir alt kümesi ve rastgele seçilen özellikler kullanılarak oluşturulur, bu da modelin genelleme yeteneğini artırır ve aşırı uyum (overfitting) riskini azaltır. Sınıflandırma problemlerinde, Random Forest her ağacın tahminini alarak en çok oyu alan sınıfı belirlerken, regresyon problemlerinde ağaçların tahminlerinin ortalamasını alır. Yüksek doğruluk, dayanıklılık ve özellik önem derecelendirmesi gibi avantajları sayesinde finans, sağlık, pazarlama ve daha birçok alanda etkin bir şekilde kullanılmaktadır. Sonuç olarak, Random Forest karmaşık veri setlerinde bile güvenilir ve doğru sonuçlar elde etmek için tercih edilen bir algoritmadır.

Random forest sınıflandırma modeli aşağıdaki Denklem 1 deki gibi hesaplanmaktadır.

$$Gini = 1 - \sum_{i=1}^C (p_i)^2$$

Gini: Gini Sıfırlığı, bir düğümdeki verilerin ne kadar homojen ya da heterojen olduğunu gösterir. Gini, saf bir düğümde (tüm örnekler aynı sınıfa ait) 0 olur. Karışıklık arttıkça Gini değeri yükselir ve maksimum karışıklık durumunda 0.5'e yaklaşır (ikili sınıflandırmada).

Σ : Toplam işareti, tüm sınıflar için birikimli hesaplama yapılır.

C: Sınıf sayısı, bir düğümdeki farklı sınıfların toplam sayısını ifade eder. Örneğin, iki sınıflı bir problemde $C = 2$ olacaktır.

p_i : i -inci sınıfın olasılığı, bir düğümdeki verilerin i -inci sınıfa ait olma oranıdır. Bu, 0 sınıfa ait veri sayısının toplam veri sayısına bölünmesiyle hesaplanır.

$(p_i)^2$: Her sınıfın olasılığının karesi alınır, çünkü Gini sıfırlık formülünde her sınıfın karesi kullanılarak hesaplanır.

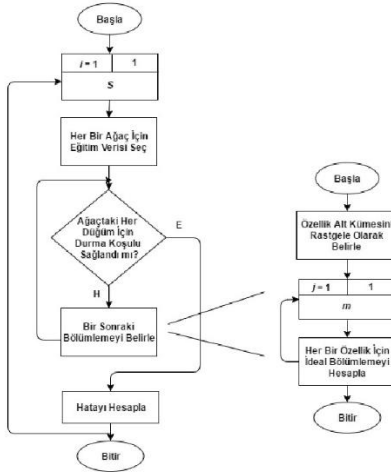
Denklem 1. Random Forest Gini Formülü

$$H(S) = - \sum_{i=1}^C p_i \log_2(p_i)$$

1. $H(S)$: Verilen S veri kümesinin entropisi.
2. C : Sınıf sayısı (veri kümesinde kaç farklı sınıf olduğu).
3. p_i : i -inci sınıfa ait veri örneklerinin oranı (olasılığı):

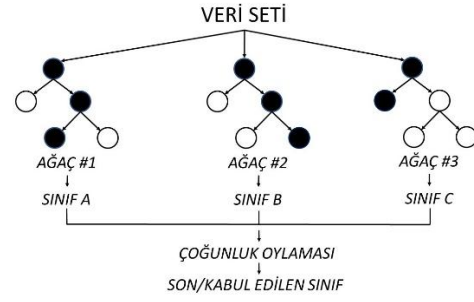
$$p_i = \frac{\text{Sınıf } i\text{'deki veri sayısı}}{\text{Toplam veri sayısı}}$$
4. $\log_2(p_i)$: p_i 'nin taban 2 logaritması (ikilik bilgi sistemi için kullanılır).

Denklem 2. Random Forest Entropi Formülü



Şekil 1. Random Forest Akış Şeması

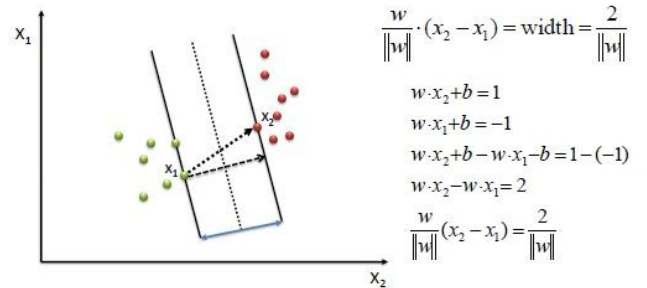
Şekil 1 deki akış şeması, Random Forest algoritmasının temel işleyişini açıklamaktadır. İlk adımda, her bir karar ağacı için eğitim verisi bootstrap örnekleme yöntemiyle seçilir. Daha sonra, her düğümde bir durma koşulu kontrol edilir; eğer koşul sağlanırsa düğüm sonlandırılır, sağlanmazsa uygun özellik alt kümesi rastgele seçilir ve düğüm bölme işlemi devam eder. Seçilen özellikler için ideal bölme noktası Gini Impurity veya Entropi gibi ölçütlere göre belirlenir. Bu süreç her bir karar ağacı için tekrarlanır ve ağacın tamamlanmasının ardından hata oranı hesaplanır. Son olarak, tüm ağaçların sonuçları birleştirilerek nihai model oluşturulur. Bu yapı, çeşitlilik sağlayarak modelin genelleme kapasitesini artırır.



Şekil 2. Random Forest Algoritması Temel İşleyiş

3.2.2 SVM

Support Vector Machines (SVM), makine öğrenmesinde hem sınıflandırma hem de regresyon problemlerini çözmek için kullanılan güçlü ve etkili bir algoritmadır. SVM, veri noktalarını farklı sınıflara ayıran bir karar sınırı (hiper düzlem) bulmayı amaçlar. Algoritmanın temel amacı, bu karar sınırını seçerken sınıflar arasındaki mesafeyi (marjin) maksimize etmektir. Şekil 1 de yöntemin uygulanma stili sunulmuştur. Bu, modeli genelleştirilebilir hale getirir ve yeni verilerle doğru tahmin yapmasını sağlar. SVM, doğrusal olarak ayrılabilen verilerde doğrusal bir hiper düzlem, doğrusal olmayan verilerde ise çekirdek (kernel) fonksiyonlarını kullanarak verileri daha yüksek boyutlu bir uzaya dönüştürerek ayrım yapar. Özellikle küçük veri setlerinde ve yüksek boyutlu verilerde etkili olmasıyla bilinen SVM, görüntü işleme, metin sınıflandırma ve biyoinformatik gibi birçok alanda yaygın olarak kullanılmaktadır.



Şekil 3. SVM genel yapısı ve formülü

3.2.3 Decision Tree

Decision Tree, makine öğrenmesinde hem sınıflandırma hem de regresyon problemlerini çözmek için kullanılan, kolay anlaşılır ve görselleştirilebilir bir algoritmadır. Veri, ağaç yapısı

şeklinde hiyerarşik olarak dallara ayrılarak işlenir. Her düğüm, bir özelliğe veya soruya dayanarak veriyi böler, yaprak düğümler ise nihai tahminleri veya sınıfları temsil eder. Bölünme kararları genellikle Gini Impurity (Denklem 1) veya Entropi (Denklem 2) gibi ölçütlerle yapılır ve hedef, veriyi en iyi şekilde ayırarak karışıklığı azaltmaktır. Decision Tree, hızlı ve yorumlanabilir olmasıyla avantaj sağlarken, aşırı uyum (overfitting) riski taşıdığı için dikkatli kullanılması gerekir. Sağlık, finans ve pazarlama gibi birçok alanda etkili bir yöntemdir.

3.2.4 Gradient Boosting

Gradient Boosting, makine öğrenmesinde güçlü bir topluluk (ensemble) öğrenme yöntemidir ve genellikle hem sınıflandırma hem de regresyon problemlerinde kullanılır. Temel olarak, zayıf öğreniciler (genellikle karar ağaçları) bir araya getirilerek güçlü bir model oluşturulur. Bu yöntem, her bir yeni ağacın önceki modelin hatalarını öğrenerek tahmin doğruluğunu artırması prensibine dayanır. "Gradient" (gradyan), kayıp fonksiyonunu en aza indirmek için hataların yönünü belirlemek amacıyla kullanılır. Gradient Boosting, yüksek doğruluk, esneklik ve güçlü genelleştirme kabiliyetiyle finans, pazarlama ve tahmin gibi alanlarda yaygın olarak kullanılan etkili bir yöntemdir.

$$F_m(x) = F_{m-1}(x) + \gamma_m h_m(x)$$

$F_m(x)$: m -inci adımda elde edilen model.

$F_{m-1}(x)$: $m - 1$ -inci adımda elde edilen model.

$h_m(x)$: m -inci adımda öğrenilen zayıf öğrenici (örneğin bir karar ağacı).

γ_m : Zayıf öğrenicinin katkısını ölçekleyen bir ağırlık katsayısı.

x : Veri noktası.

Denklem 3. Gradient Boosting Formülü

3.2.5 Logistic Regression

Logistic Regression, makine öğrenmesinde sınıflandırma problemleri için kullanılan temel ve etkili bir algoritmadır. Adından "regresyon" geçmesine rağmen, doğrusal regresyonun aksine sürekli bir değer tahmini yerine sınıflar arasında bir ayırım yapmayı amaçlar. Logistic Regression, bir veri noktasının belirli bir sınıfa ait olma olasılığını

tahmin etmek için sigmoid (lojistik) fonksiyonunu kullanır. Bu sayede tahmin edilen değerler 0 ile 1 arasında sınırlanır ve bir eşik değeri belirlenerek sınıflandırma yapılır. Algoritma, özellikle ikili sınıflandırma problemleri için uygundur ve finans, sağlık ve sosyal bilimlerde gibi birçok alanda yaygın olarak kullanılmaktadır.

$$h_\theta(x) = \frac{1}{1 + e^{-(\theta^T x)}}$$

$h_\theta(x)$: Tahmin edilen olasılık ($P(y = 1|x)$).

$\theta^T x$: Veri ve ağırlık vektörünün iç çarpımı (doğrusal kombinasyon).

$$\theta^T x = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \dots + \theta_n x_n$$

- θ : Model parametreleri (ağırlıklar).
- x : Giriş özellik vektörü.

e : Doğal logaritmanın tabanı (Euler sayısı, yaklaşık 2.718).

Denklem 4. Logistic Regression Formülü

3.2.6 Gaussian NB

Gaussian Naive Bayes (Gaussian NB), makine öğrenmesinde yaygın olarak kullanılan ve özellikle sınıflandırma problemleri için etkili olan basit ve hızlı bir algoritmadır. Naive Bayes sınıflandırıcısının bir türüdür ve sürekli özelliklere sahip veri kümeleriyle çalışırken özelliklerin normal dağılıma (Gaussian dağılımı) sahip olduğunu varsayar. Algoritma, Bayes Teoremi'ni temel alır ve her özelliğin bağımsız olduğunu varsayar (bu nedenle "Naive" denir). Gaussian NB, sınıfları belirlemek için her bir özelliğin olasılık yoğunluk fonksiyonunu hesaplar ve maksimum olasılığa sahip sınıfı tahmin eder. Algoritmanın basit yapısı ve düşük hesaplama maliyeti sayesinde metin sınıflandırma, spam tespiti ve genetik analiz gibi birçok alanda kullanılır.

$$P(C|X) = \frac{P(X|C) \cdot P(C)}{P(X)}$$

$P(C|X)$: X gözleminin C sınıfına ait olma olasılığı (posterior probability).

$P(X|C)$: X gözleminin C sınıfı altında gerçekleşme olasılığı (likelihood).

$P(C)$: C sınıfının önceden bilinen olasılığı (prior probability).

$P(X)$: X gözleminin genel olasılığı (normalizasyon faktörü).

Denklem 5. Gaussian NB Formülü

3.2.7 K-Nearest Neighbors (KNN)

K-Nearest Neighbors (KNN), hem sınıflandırma hem de regresyon problemlerinde kullanılan basit ve etkili bir makine öğrenmesi algoritmasıdır. KNN, yeni bir veri noktası için tahmin yaparken, veri setindeki en yakın KKK komşuyu bulur ve bu komşuların etiketlerine veya değerlerine dayanarak tahminde bulunur. Sınıflandırma durumunda, çoğunluk oyu prensibiyle tahmin yapılır; regresyonda ise komşu değerlerin ortalaması alınır. KNN, özellikler arasındaki benzerliği ölçmek için genellikle **Öklid mesafesi** kullanır ve bu nedenle algoritmanın başarısı, verilerin ölçeklendirilmesine ve uygun KKK değerinin seçilmesine bağlıdır. Basit yapısı ve açıklanabilirliği nedeniyle KNN, metin sınıflandırma, öneri sistemleri ve görüntü işleme gibi birçok alanda tercih edilir.

$$d(x, x_i) = \sqrt{\sum_{j=1}^n (x_j - x_{ij})^2}$$

$d(x, x_i)$: Tahmin yapılacak x noktası ile veri setindeki x_i noktası arasındaki mesafe.

n : Özellik (boyut) sayısı.

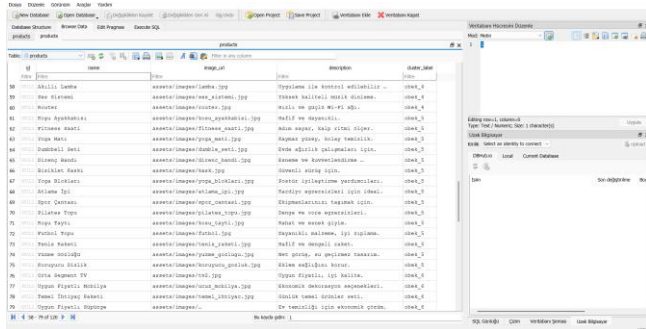
x_j : x noktası için j -inci özellik değeri.

x_{ij} : x_i noktası için j -inci özellik değeri.

Denklem 6. KNN Formülü

3.3 SQLite Veritabanı

Proje kapsamında verilerin saklanması ve yönetilmesi için SQLite veritabanı tercih edilmiştir. Bu sayede, raporlama aşamasında ihtiyaç duyulan bilgiler tek bir dosya içerisinde düzenli bir biçimde tutulmuş, sorgular hızlı ve verimli bir şekilde gerçekleştirilmiştir. Ayrıca, SQLite'in basit kurulumu ve bakım kolaylığı, veri işleme süreçlerini yalınlaştırarak raporların oluşturulma süresini kısaltmıştır. Uygulama üzerinde bulunan reklam önerme fonksiyonu için gerekli olan verileri veritabanı üzerinden sağlıyoruz.



id	name	description	category
30	Abdül Sami	ayıklama ile kemeri süslüleştirir...	örnek_3
31	Abdül Sami	ayıklama ile kemeri süslüleştirir...	örnek_3
32	Abdül Sami	ayıklama ile kemeri süslüleştirir...	örnek_3
33	Abdül Sami	ayıklama ile kemeri süslüleştirir...	örnek_3
34	Abdül Sami	ayıklama ile kemeri süslüleştirir...	örnek_3
35	Abdül Sami	ayıklama ile kemeri süslüleştirir...	örnek_3
36	Abdül Sami	ayıklama ile kemeri süslüleştirir...	örnek_3
37	Abdül Sami	ayıklama ile kemeri süslüleştirir...	örnek_3
38	Abdül Sami	ayıklama ile kemeri süslüleştirir...	örnek_3
39	Abdül Sami	ayıklama ile kemeri süslüleştirir...	örnek_3
40	Abdül Sami	ayıklama ile kemeri süslüleştirir...	örnek_3
41	Abdül Sami	ayıklama ile kemeri süslüleştirir...	örnek_3
42	Abdül Sami	ayıklama ile kemeri süslüleştirir...	örnek_3
43	Abdül Sami	ayıklama ile kemeri süslüleştirir...	örnek_3
44	Abdül Sami	ayıklama ile kemeri süslüleştirir...	örnek_3
45	Abdül Sami	ayıklama ile kemeri süslüleştirir...	örnek_3
46	Abdül Sami	ayıklama ile kemeri süslüleştirir...	örnek_3
47	Abdül Sami	ayıklama ile kemeri süslüleştirir...	örnek_3
48	Abdül Sami	ayıklama ile kemeri süslüleştirir...	örnek_3
49	Abdül Sami	ayıklama ile kemeri süslüleştirir...	örnek_3
50	Abdül Sami	ayıklama ile kemeri süslüleştirir...	örnek_3
51	Abdül Sami	ayıklama ile kemeri süslüleştirir...	örnek_3
52	Abdül Sami	ayıklama ile kemeri süslüleştirir...	örnek_3
53	Abdül Sami	ayıklama ile kemeri süslüleştirir...	örnek_3
54	Abdül Sami	ayıklama ile kemeri süslüleştirir...	örnek_3
55	Abdül Sami	ayıklama ile kemeri süslüleştirir...	örnek_3
56	Abdül Sami	ayıklama ile kemeri süslüleştirir...	örnek_3
57	Abdül Sami	ayıklama ile kemeri süslüleştirir...	örnek_3
58	Abdül Sami	ayıklama ile kemeri süslüleştirir...	örnek_3
59	Abdül Sami	ayıklama ile kemeri süslüleştirir...	örnek_3
60	Abdül Sami	ayıklama ile kemeri süslüleştirir...	örnek_3

Uygulama 1. SQL Veritabanı Gorseli

3.4 Streamlit Arayuzu

Proje kapsamında, veri görselleştirme ve kullanıcı dostu arayüz sağlamak amacıyla **Streamlit** kullanılmıştır. Bu sayede, raporlamaya yönelik veriler sade bir web arayüzü üzerinden dinamik olarak sunulmuş, filtreleme ve etkileşimli grafikler sayesinde kullanıcıların veriye hızlıca erişmesi ve içgörü elde etmesi kolaylaştırılmıştır. Ayrıca, **CSS fonksiyonları** ile arayüz tasarımı özelleştirilerek daha estetik, kurumsal bir görünüm ve kullanıcı deneyimi sağlanmıştır.

Reklam Stratejisi için Öbek Tabanlı Tahmin Uygulaması

Bu uygulama, girilen kullanıcı verilerine dayalı olarak öbek tahmini yapar, reklam stratejisi ve ürün önerisi sunar. Ayrıca yeni tahminleri veri setine ekleyerek modeli yeniden eğitmeye imkan tanır.

Kullanıcı Verilerini Girin

Cinsiyet

Kadın

Yaş Grubu

18-30

Medeni Durum

Evli

Eğitim Düzeyi

Eğitimsiz

İstihdam Durumu

İşsiz veya Düzenli Bir İş Yok

Yıllık Ortalama Gelir (TL)

0,00

Yaşadığı Şehir

Büyük Şehir

Yıllık Ortalama Satın Alım Miktarı (TL)

Uygulama 2. Streamlit Arayüz Gorseli

3.5 Modelin Yeniden Eğitilmesi

Kullanıcıdan alınan yeni verilerin data/new_data.csv dosyasına eklenmesini ve bu eklenen verilerle ana eğitim verisinin (train.csv) birleştirilerek modelin yeniden eğitilmesini sağlamaktadır. Böylece, model zaman içinde gelen yeni bilgilere uyum sağlayarak güncel ve daha doğru sonuçlar üretir. Bu yaklaşım, raporlama süreçlerine eklenen verilerle modelin sürekli geliştirilebilmesini ve güncellenen verilere dayalı içgörülerin elde edilmesini mümkün kılar.

Modeli Yeni Verilerle Yeniden Eğit

Bu butona basıldığında 'train.csv' ile 'new_data.csv' birleştirilerek model yeniden eğitilir.

Yeniden Eğit

Uygulama 3. Yeniden Eğitim Gorseli

3.6 Reklam Onerisi

Uygulamada, kullanıcı tarafından girilen veriler modele alınıp tahmini bir “öbek” (müşteri segmenti) belirlenmektedir. Bu öbek belirlendikten sonra, ilgili öbek için önceden tanımlanmış bir reklam stratejisi kullanıcıya sunulur. Örneğin, obek_1 için “İhtiyaç Odaklı Yaşayanlar” stratejisi gösterilirken, obek_4 için “Bolluk İçinde Yaşayanlar” denilmektedir. Bu sayede, kullanıcıya sunulan reklam veya ürün önerileri, tahmini öbeğe özel olarak şekillendirilmiş, hedefli bir pazarlama yaklaşımı sağlamaktadır. Bu reklamları 3.3 de daha önceden kaydettiğimiz ürünler veritabanından çekiyoruz.

Tahmin Edilen Öbek: obek_1

Onerilen Reklam Stratejisi: Eğitim odaklı ürünler önerilir.

Bu Öbeğe Özel Ürün Önerileri



Seminer

Kariyer semineri kaydı.



Gelişim Kitabı

Kişisel gelişim kitabı.



Liderlik Eğitimi

Liderlik ve takım yönetimi eğitimi.

Yeni tahmin verisetine kaydedildi.

Uygulama 4. Reklam Onerisi Gorseli

4 Uygulama

4.1 Veri Kümesi

Bu çalışmada kullanılan iki veri seti, bireylerin demografik ve tüketim alışkanlıklarını içermektedir. Eğitim veri seti (train.csv), 5460 gözlem ve 14 özellikten oluşurken, test veri seti (test_x.csv) 2340 gözlem ve 13 özellik içermektedir. Her iki veri setinde de cinsiyet, yaş grubu, medeni durum, eğitim düzeyi, istihdam durumu, yıllık gelir gibi demografik bilgiler ile en çok ilgilenilen ürün grubu, yıllık satın alım miktarı gibi tüketim alışkanlıklarını yansıtan bilgiler bulunmaktadır. Eğitim veri setinde ayrıca bireylerin öbek isimleri (hedef değişken) yer alırken, test veri setinde bu bilgi bulunmamaktadır. Eksiksiz ve tam olan bu veri

setleri, bireylerin öbeklere atanmasını tahmin etmeye yönelik bir model oluşturmak için kullanılacaktır.

Değişken Adı	Açıklama	Örnek Veri
Cinsiyet	Bireyin cinsiyeti	Erkek, Kadın
Yaş Grubu	Bireyin ait olduğu yaş grubu	18-30, 31-40, 41-50, 51-60, >60
Medeni Durum	Evli veya bekar durumu	Bekar, Evli
Eğitim Düzeyi	Bireyin eğitim düzeyi	Lisans, Yüksek Lisans, Lise, vb.
İstihdam Durumu	Bireyin istihdam durumu	Çalışıyor, İşsiz, Serbest Meslek, Emekli
Yıllık Ortalama Gelir	Bireyin yıllık ortalama geliri	74826; 246298; ...
Yaşadığı Şehir	Bireyin yaşadığı şehir türü	Büyük Şehir, Küçük Şehir, Kırsal, Köy
En Çok İlgilendiği Ürün Kategorisi	Bireyin en çok ilgilendiği ürün kategorisi	Elektronik, Giyim, Spor Ekipmanları, vb.
Yıllık Ortalama Satın Alım Miktarı	Bireyin yıllık ortalama satın aldığı ürün miktarı	360; 120; ...
Yıllık Ortalama Sipariş Verilen Ürün Adedi	Bireyin yıllık ortalama sipariş verdiği ürün sayısı	-
Eğitime Devam Etme Durumu	Bireyin eğitime devam edip etmediği	Devam Ediyor, Devam Etmiyor
Yıllık Ortalama Sepete Atılan Ürün Adedi	Bireyin yıllık ortalama sepete eklediği ürün adedi	Y, N
Sınıflandırma Etiketleri	Sınıflandırma modeli ile tahmin edilmesi beklenen etiket veya kategori	obek_1, obek_2, obek_3, ...

Tablo 1. Veriseti Oznitelikler Tablosu

Tablo 1’de , çalışmada kullanılan veri setindeki değişkenlerin adlarını, açıklamalarını ve örnek veri içeriklerini özetlemektedir. Değişkenler arasında bireylerin demografik bilgileri (cinsiyet, yaş grubu, medeni durum, eğitim düzeyi, istihdam durumu), coğrafi konumu (yaşadığı şehir), gelir durumu (yıllık ortalama gelir) ve tüketim alışkanlıklarını (en çok ilgilendiği ürün kategorisi, yıllık satın alım miktarı, sipariş verilen ürün adedi) açıklayan bilgiler bulunmaktadır. Ayrıca, bireylerin eğitim durumu ve sınıflandırma etiketi gibi modelleme sürecinde kullanılan hedef değişkenler de yer almaktadır. Bu bilgiler, bireylerin sınıflandırılması için temel veri kaynağını oluşturur.

```
train.duplicated().sum()
[8]
0
```

```
train.isnull().sum()
[7]
Cinsiyet 0
Yaş Grubu 0
Medeni Durum 0
Eğitim Düzeyi 0
İstihdam Durumu 0
Yıllık Ortalama Gelir 0
Yaşadığı Şehir 0
En Çok İlgilendiği Ürün Grubu 0
Yıllık Ortalama Satın Alım Miktarı 0
Yıllık Ortalama Sipariş Verilen Ürün Adedi 0
Eğitime Devam Etme Durumu 0
Öbek İsmi 0
Yıllık Ortalama Sepete Atılan Ürün Adedi 0
dtype: int64
```

Şekil 4. Verisetinde bos veri ve tekrar eden veri kontrol edildi

	Yıllık Ortalama Gelir	Yıllık Ortalama Satın Alım Miktarı	Yıllık Ortalama Sipariş Verilen Ürün Adedi	Yıllık Ortalama Sepete Atılan Ürün Adedi
count	5.460000e+03	5460.000000	5460.000000	5460.000000
mean	3.635711e+05	16616.612217	24.040884	73.445693
std	2.197144e+05	14099.171704	14.945655	47.214184
min	4.392299e+04	2859.254000	0.000000	3.977559
25%	2.156934e+05	4931.859057	11.550502	25.009168
50%	2.869254e+05	8426.818967	20.095924	82.485579
75%	4.681882e+05	30579.244695	35.918161	104.473291
max	1.192437e+06	48605.594415	64.616196	242.308441

Tablo 2. Sürekli Değerlerin İstatistiksel Sonuçları

Tablo 2 de, veri setindeki sürekli değişkenlerin temel istatistiksel özetlerini sunmaktadır. Değişkenler arasında Yıllık Ortalama Gelir, Yıllık Ortalama Satın Alım Miktarı, Yıllık Ortalama Sipariş Verilen Ürün Adedi ve Yıllık Ortalama Sepete Atılan Ürün Adedi yer almaktadır. Her bir değişken için toplam gözlem sayısı (count), ortalama (mean), standart sapma (std), minimum değer (min), çeyreklik değerler (25%, 50%, 75%) ve maksimum değer (max) hesaplanmıştır. Bu özet, değişkenlerin dağılımını ve verinin genel yapısını anlamaya yardımcı olmaktadır.

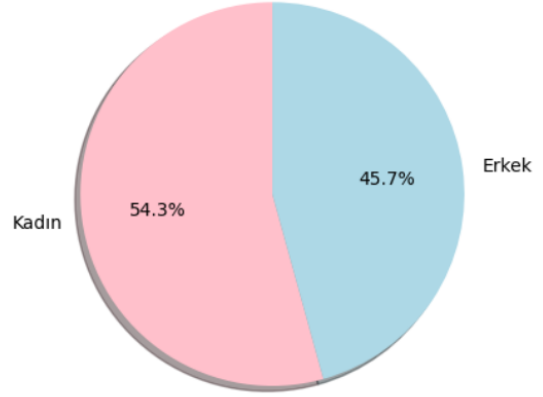
$$\text{Mean} = \frac{\sum_{i=1}^n x_i}{n} \quad \text{Std} = \sqrt{\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n - 1}}$$

$$\text{IQR} = Q3 - Q1 \quad \begin{array}{l} \text{Alt sınır: } Q1 - 1.5 \times IQR \\ \text{Üst sınır: } Q3 + 1.5 \times IQR \end{array}$$

Denklem 7. İstatistiksel Formüller

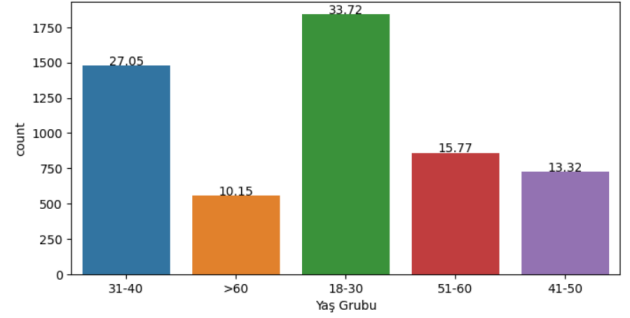
4.1.1 Veri Kümesi Kategorik Verilerin Grafikleri

Simdi verisetindeki kategorik verilerin dağılımlarını gösteren grafikleri inceleyelim.



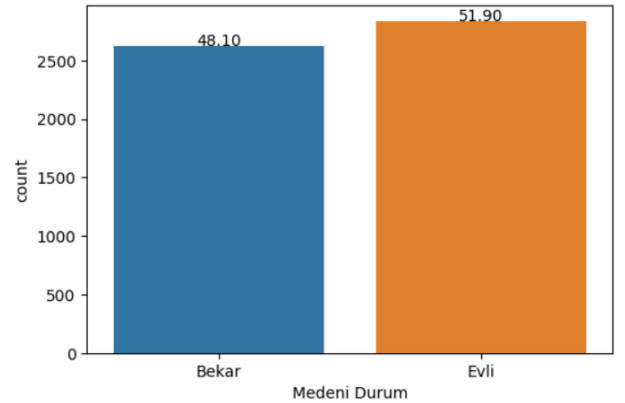
Şekil 5. Kadın Erkek Dağılımı

Şekil 5'deki grafikte açıkça görebiliyoruz ki, veri setinde Kadınların yüzdesi %54.3, Erkeklerin yüzdesi ise %45.7'dir.



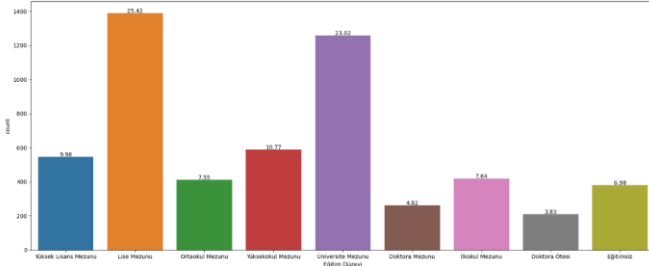
Şekil 6. Yaş grubu Dağılımı

Şekil 6'da görüldüğü gibi 18 ile 30 yaş arasında olan kişilerin sayısı diğer yaş gruplarına kıyasla daha fazladır.



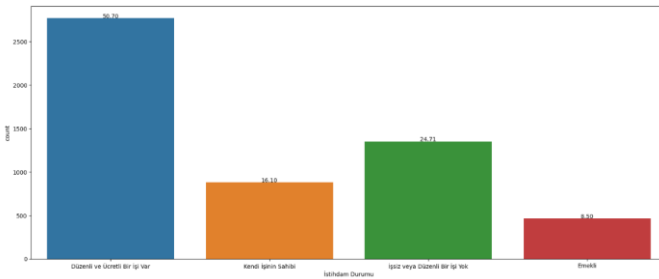
Şekil 7. Medeni Durum Tablosu

Şekil 7'deki grafikte görüldüğü gibi evli olan kişilerin sayısı, bekâr olanlara kıyasla daha fazladır.



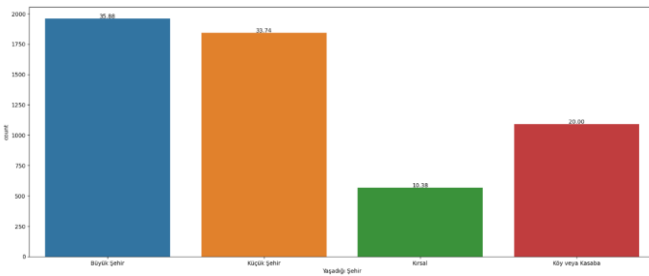
Şekil 8. Eğitim Durum Dağılımı

Şekil 8'de görüldüğü üzere liseden ve üniversiteden mezun olan kişilerin sayısı, diğer eğitim seviyelerine göre daha fazladır.



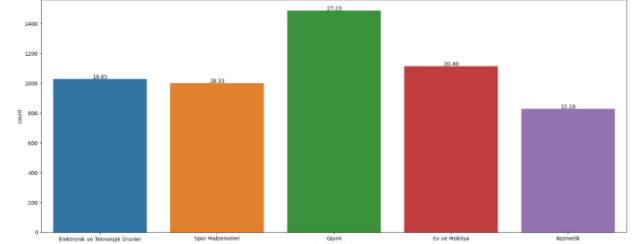
Şekil 9. İstihdam Durumu Tablosu

Şekil 9'da görüldüğü üzere düzenli ve ücretli bir işi olan kişilerin sayısı oldukça fazladır. İşsiz olan ve düzenli bir işi olmayanların oranı ise %24.71'dir.



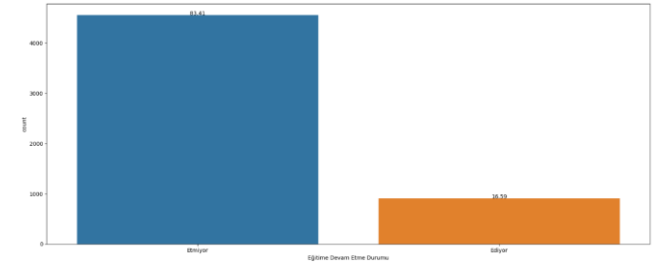
Şekil 10. Tasadığı Şehir Tablosu

Şekil 10'da görüldüğü üzere büyük şehirlerde ve küçük şehirlerde yaşayan insanların sayısı fazladır.

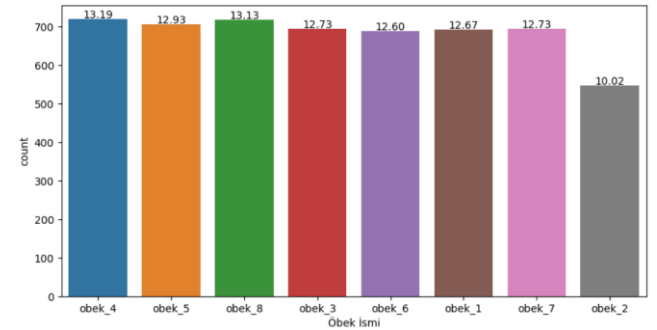


Şekil 11. Kullanıcıların En Çok İlgilendikleri Ürünler

Şekil 11'de görüldüğü üzere en çok ilgi duyulan giyim ürünleri oldu



Şekil 12. Eğitime Devam Etme Durumu

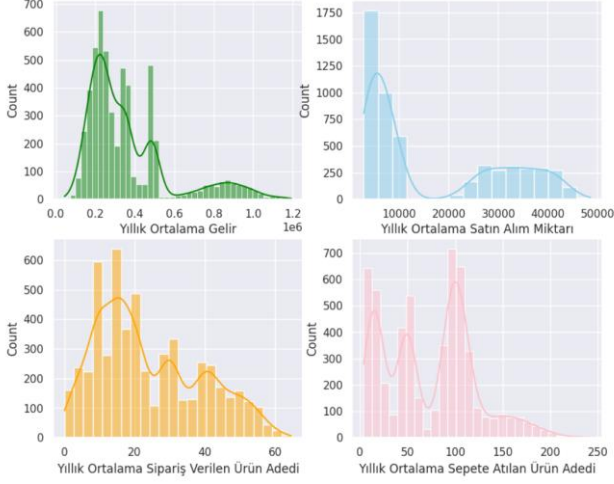


Şekil 13. Obek Dağılımı

Şekil 13 de görüldüğü gibi 8 farklı sınıf etiketi bulunmaktadır ve bu etiketler dengesiz bir dağılıma sahiptir. Problemimiz, dengesiz çok sınıflı sınıflandırma problemidir.

4.1.2 Veri Kümesi Surekli Verilerin Grafikleri

Simdi verisetindeki surekli verilerin dagilimlarini gosteren grafikleri inceleyelim.



Şekil 14. Surekli verilerin grafikleri

Sekil 14'deki Grafikler, yıllık ortalama gelir, yıllık ortalama satın alım miktarı, yıllık ortalama sipariş verilen ürün sayısı ve yıllık ortalama sepete eklenen ürün sayısının yoğunluk dağılımını göstermektedir. Sol üst köşede, yıllık ortalama gelire baktığımızda, gelir dağılımının sola çarpık olduğu ve veri içinde 3 farklı normal dağılımın bulunduğu görülmektedir. Sağ üst köşede, yıllık ortalama satın alım miktarının 2 farklı dağılıma sahip olduğu görülmektedir. Sol alt köşede birçok normal dağılım bulunmaktadır. Sağ alt köşede ise 3 farklı normal dağılım bulunmaktadır. Bu 4 grafiğin genel olarak verdiği fikir, çok değişkenli bir normal dağılımın var olduğudur. Bu nedenle, veriyi normalize etmemiz gerekmektedir.

4.1.3 Diğer Analizler

Bu bölüm, her bir değişken için ek keşifleri gösterecektir. Ek keşifler şunlardır:

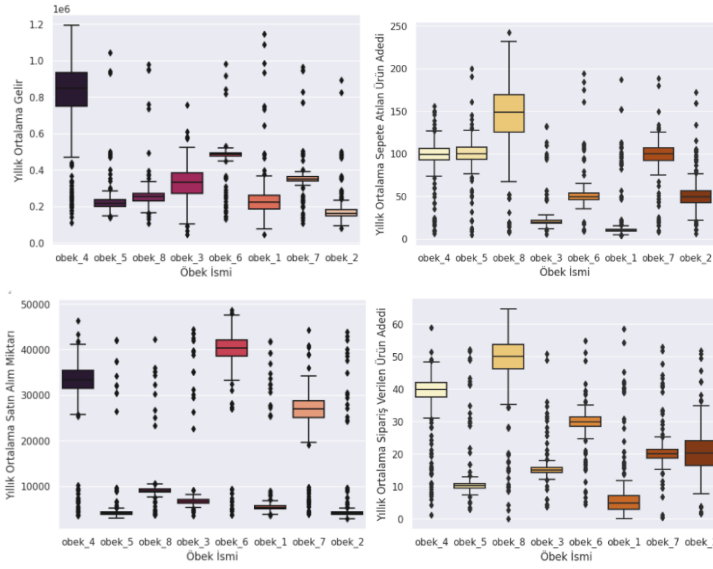
Isı Haritası (Heatmap)

Kategorik - Sayısal Veri Arasındaki ilişki



Şekil 15. Heat Map

Şekil 15'de bahsedilen Isı haritası (Heatmap), iki özellik arasındaki ilişkiyi gösterir ve veriyi analiz etmeye yardımcı olur. Eğer değer 1'e yakınsa, iki özelliğin pozitif bir ilişkiye sahip olduğunu söyleyebiliriz. Değer -1'e yakınsa, iki özellik arasında negatif bir ilişki vardır. Eğer değer 0'a eşitse, özellikler arasında bir ilişki bulunmamaktadır. Isı haritasına baktığımızda, Yıllık Ortalama Gelir ile Yıllık Ortalama Satın Alım Miktarı arasında pozitif bir korelasyon olduğunu açıkça görebiliriz.

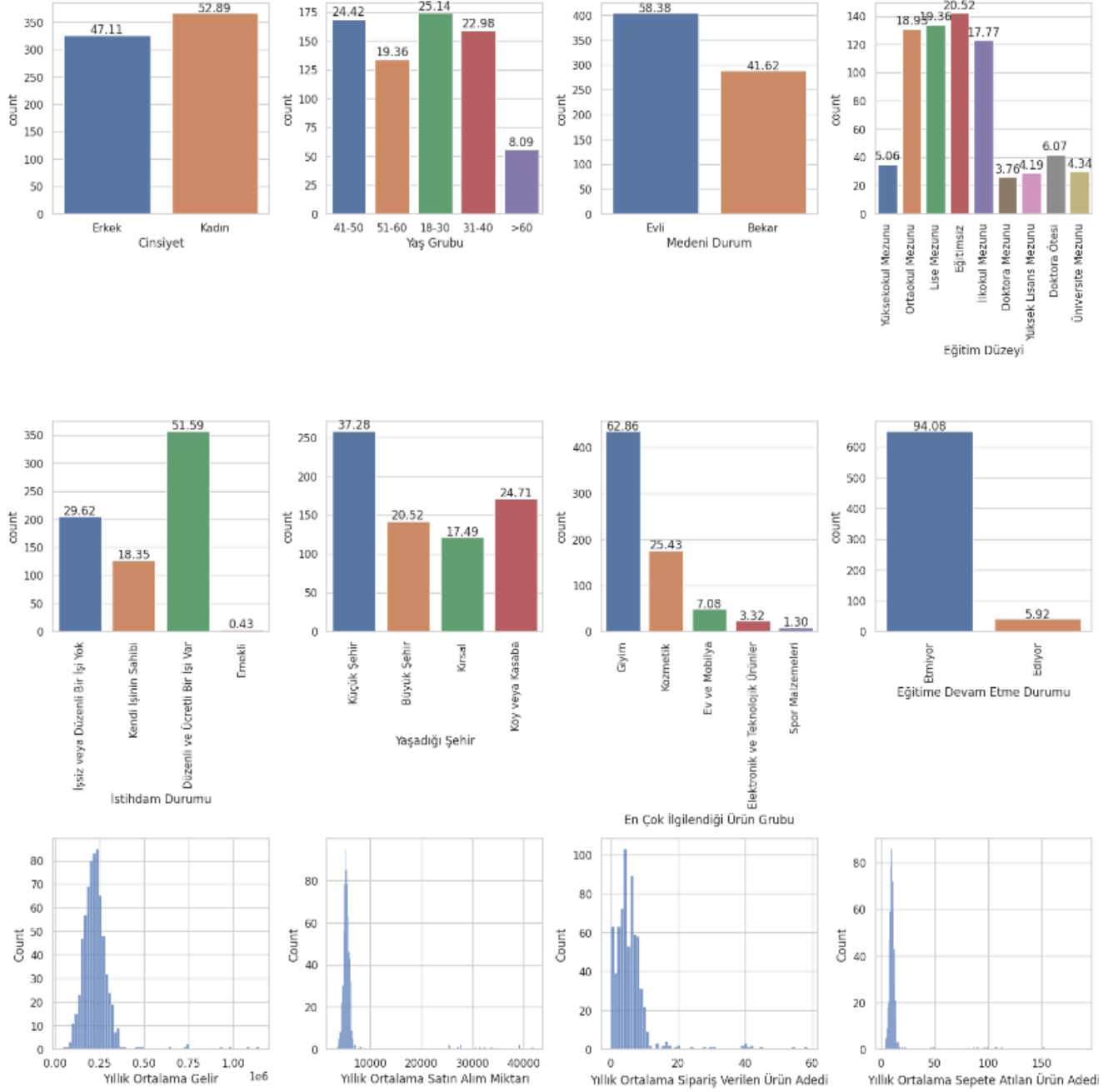


Şekil 16. Sayısal Değerlerin Kutu Grafikleri

Sekil 16’da verisetimize ait dört farklı metrik için, çeşitli “öbek” grupları arasında yapılan karşılaştırmalar yer alıyor. Her bir kutu grafiği (boxplot) belirli bir metriğin öbeklere göre dağılımını göstermekte. Soldaki üst grafik “Yıllık Ortalama Gelir” değerlerini, sağdaki üst grafik “Yıllık Ortalama Satışta Alınan Ürün Adedi”ni, soldaki alt grafik “Yıllık Ortalama Satın Alım Miktarı”ni, sağdaki alt grafik ise “Yıllık Ortalama Sipariş Verilen Ürün Adedi”ni karşılaştırıyor. Her bir grafik x ekseninde farklı öbek isimlerine (obek_4, obek_5, obek_8, obek_3, obek_6, obek_1, obek_7, obek_2) göre ayrılmış olup, dikey ekseninde ilgili metriklerin dağılımı, medyan değeri, çeyreklikleri ve uç değerleri gözlemlenebiliyor. Bu sayede, her bir öbeğin yıllık gelir, satın alma ve sipariş davranışları açısından nasıl farklılaştığı, hangi grupların daha yüksek ya da daha düşük ortalama değerlere sahip olduğu ve istatistiksel olarak nasıl bir varyasyonun mevcut olduğu görsel olarak anlaşılabilir.

4.1.4 Obeklerin Analizi

Şimdi de uygulamamızda yer alan öbeklerin, tanımlı nitelikler üzerindeki dağılımlarını gösteren grafiklerin yanı sıra, bu öbeklerin kapsamlı analizlerinin yer aldığı bir bölüm olacaktır.

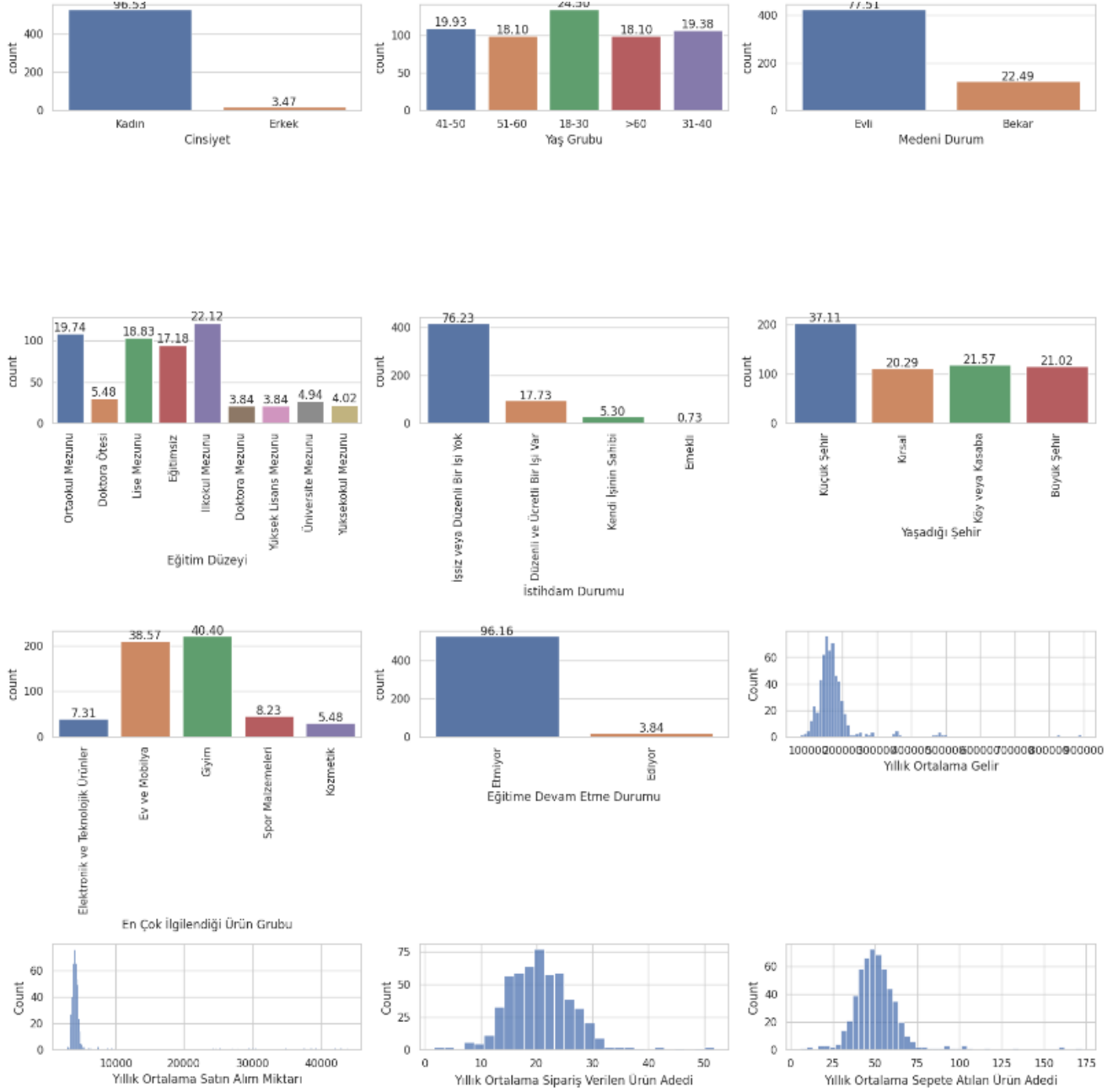


Şekil 17. Obek 1 İstatistikleri

Bu grup %47.11 erkek ve %52.89 kadından oluşmaktadır. Tüm yaş gruplarından bireyler bulunmaktadır. 18-30 yaş aralığı %25.14, 41-50 yaş aralığı %24.42, 31-40 yaş aralığı %22.98, 51-60 yaş aralığı %19.36, ve 60 yaş üstü %8.09 oranında grubu oluşturmaktadır.

Bu grup ağırlıklı olarak evli bireylerden (%58.38) ve daha düşük eğitim seviyesine sahip kişilerden oluşmaktadır. Düzenli bir işe sahip bireylerin sayısı yüksek olmakla birlikte, işsizlik oranı da dikkat çekicidir. Bu grubun ortalama gelir seviyesi diğer gruplara kıyasla daha düşüktür. Genellikle sınırlı sayıda ürünü sepete eklemekte ve sipariş etmektedirler. En çok tercih ettikleri ürün kategorileri giyim ve kozmetiktir.

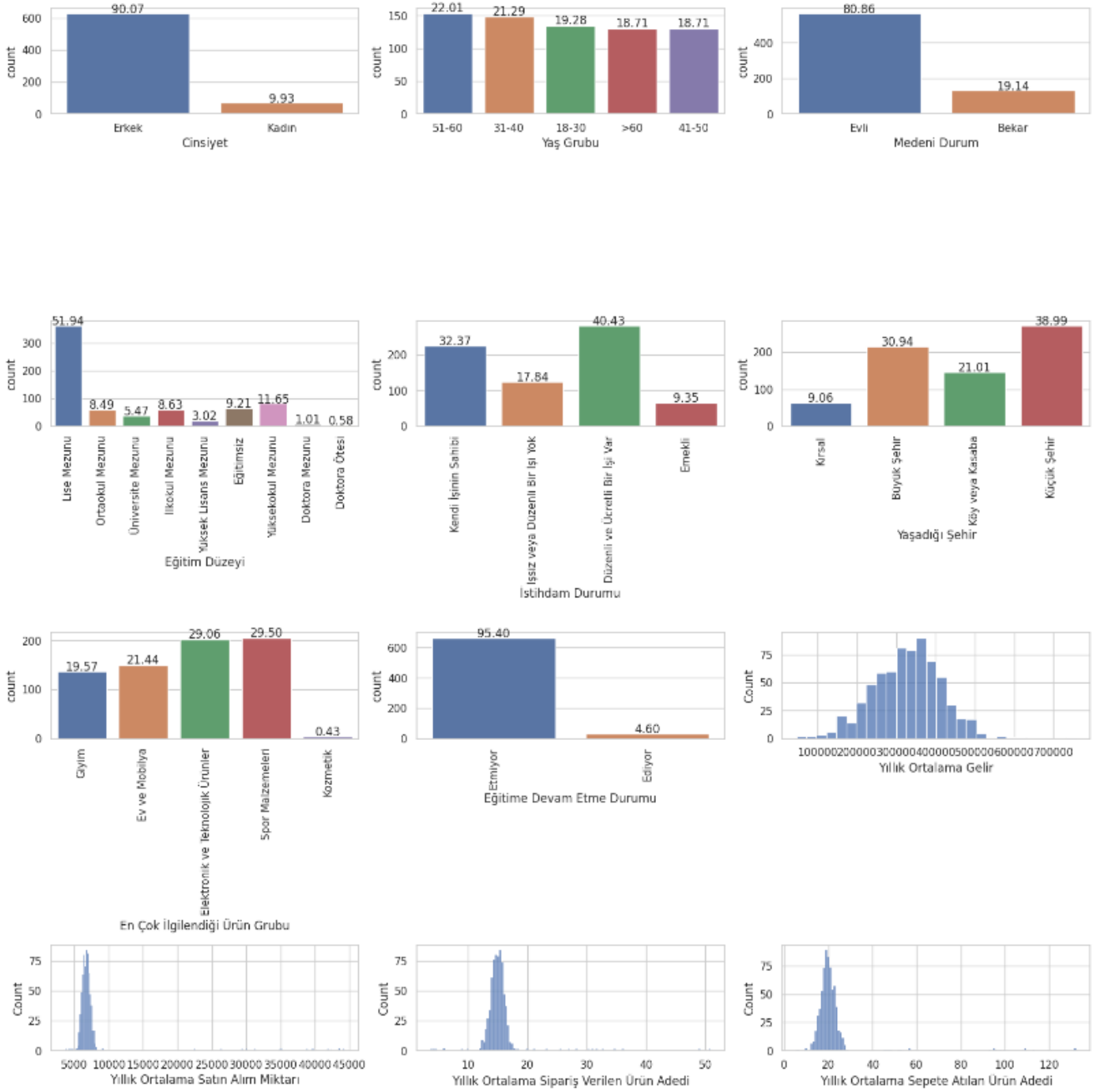
Gelir seviyelerinin sınırlı olması nedeniyle, bu gruptaki bireyler daha çok temel ihtiyaçlara odaklanmaktadır. Bu durum, sadece ihtiyaç duydukları ürünleri satın alma eğiliminde olmalarından dolayı satın alım miktarlarını azaltmaktadır. Bu bağlamda, bu grup "İhtiyaç Odaklı Yaşayanlar" olarak adlandırılabilir.



Şekil 18. Obek 2 İstatistikleri

Bu grubu oluşturanların çoğunluğu kadınlardan oluşmaktadır ve kadın oranı %96.53'e ulaşmaktadır. Bu kadınların büyük bir kısmı (%77.51) evlidir. Eğitim seviyeleri genellikle düşüktür ve istihdam edilmemişlerdir; işsizlik oranı yüksektir. Grubun %37.11'ini küçük kasabalarda yaşayan bireyler oluşturmaktadır. En çok ilgi gösterdikleri ürün kategorisi %40.40 ile giyimdir, bunu %38.57 ile ev ve mobilya ürünleri takip etmektedir.

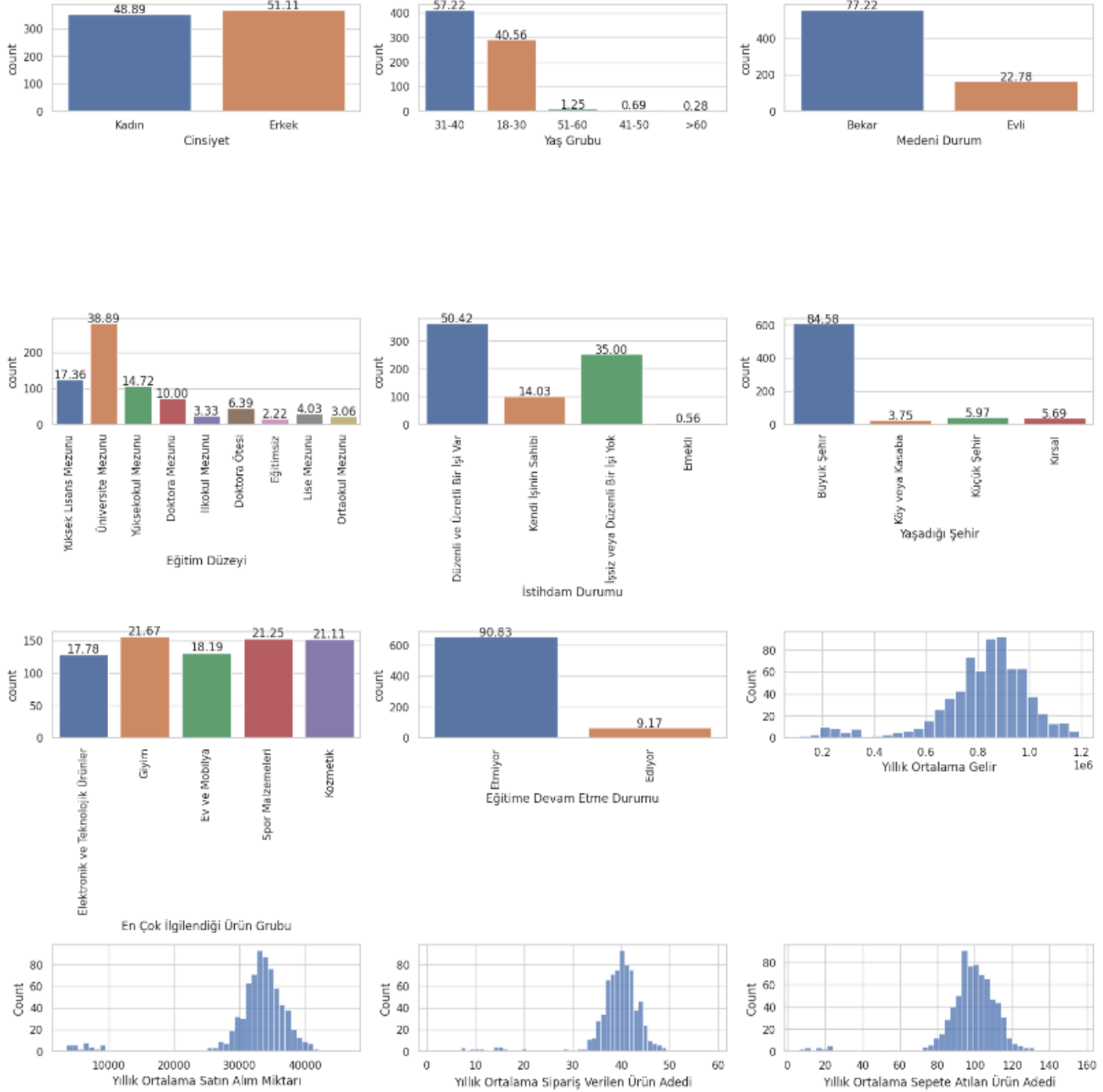
Bu grubun yıllık ortalama gelir seviyeleri ve satın alma miktarları oldukça sınırlıdır. Bu grubu diğerlerinden ayıran belirgin bir özellik, çoğunluğunun eğitimsiz ve işsiz kadınlardan oluşmasıdır. Bu nedenle, bu grup, genellikle toplum tarafından göz ardı edilen ancak ev işlerine önemli katkılarda bulunan kadınlara bir saygı duruşu niteliğinde, "Görünmez Emekçiler" olarak adlandırılabilir.



Şekil 19. Obek 3 İstatistikleri

Bu grubu oluşturan bireylerin %90.07'si erkeklerden oluşmaktadır. Farklı yaş gruplarını kapsayan bir dağılım söz konusudur. Bu grup içinde bireylerin %80.86'sı evlidir, bu da yüksek bir evlilik oranını göstermektedir. Eğitim düzeyine baktığımızda, katılımcıların %51.94'ü lise mezunudur. İstihdam durumuna göre, bu grubun önemli bir kısmı düzenli ve maaşlı işlerde çalışan bireylerden oluşmaktadır. Bu topluluğun en çok ilgilendiği alanlar spor malzemeleri, elektronik ve teknolojik ürünlerdir.

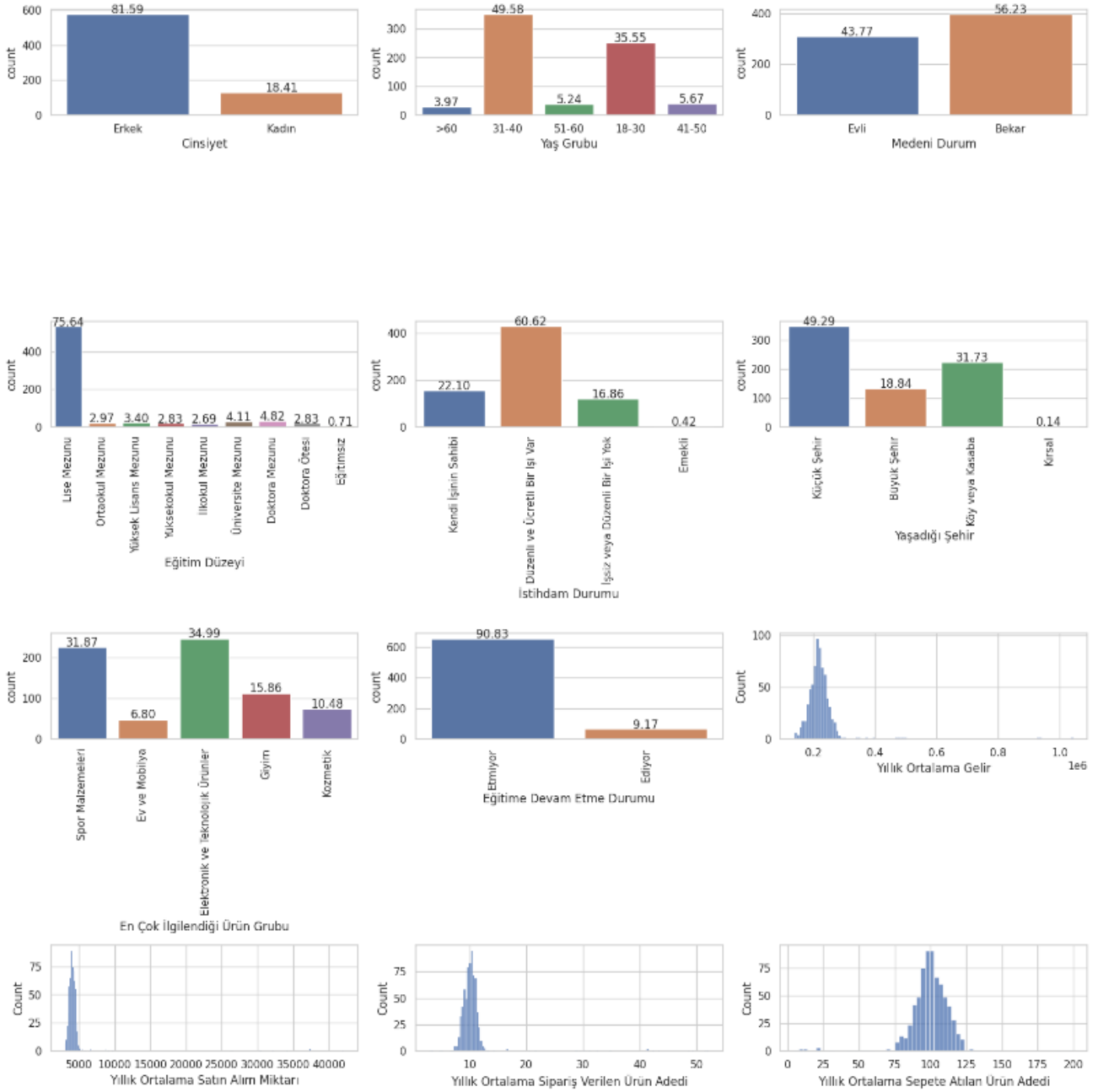
Yıllık ortalama gelirleri yüksek olmasına rağmen, bu gruptaki bireylerin satın alma değerleri düşük seviyededir. Sipariş edilen ürün sayısı ile sepete eklenen ürün sayısı birbirine yakın oranlardadır. Bu verilere dayanarak, bu grubun en belirgin özelliği evli erkeklerden oluşmasıdır ve gelirlerine rağmen satın alma miktarlarının sınırlı olmasıdır. Bu grubu "Finansal Denge Ustaları" olarak adlandırıyorum, çünkü evlilik ve ekonomik istikrarı göz önünde bulundurarak dengeli bir yaşam tarzı benimsediklerini düşünüyorum.



Şekil 20. Obek 4 İstatistikleri

Bu grubu oluşturan bireylerin %48.89'u kadın, %51.11'i erkektir. Yaş grubu genellikle 31-40 aralığında yoğunlaşmıştır ve bu grup içinde bireylerin %77.22'si bekardır. Katılımcıların eğitim düzeyi üniversite mezunu olarak belirlenmiştir. İstihdam durumuna baktığımızda, bireylerin %50.42'sinin düzenli ve maaşlı işlerde çalıştığı, ancak işsizlik oranının da yüksek olduğu görülmektedir.

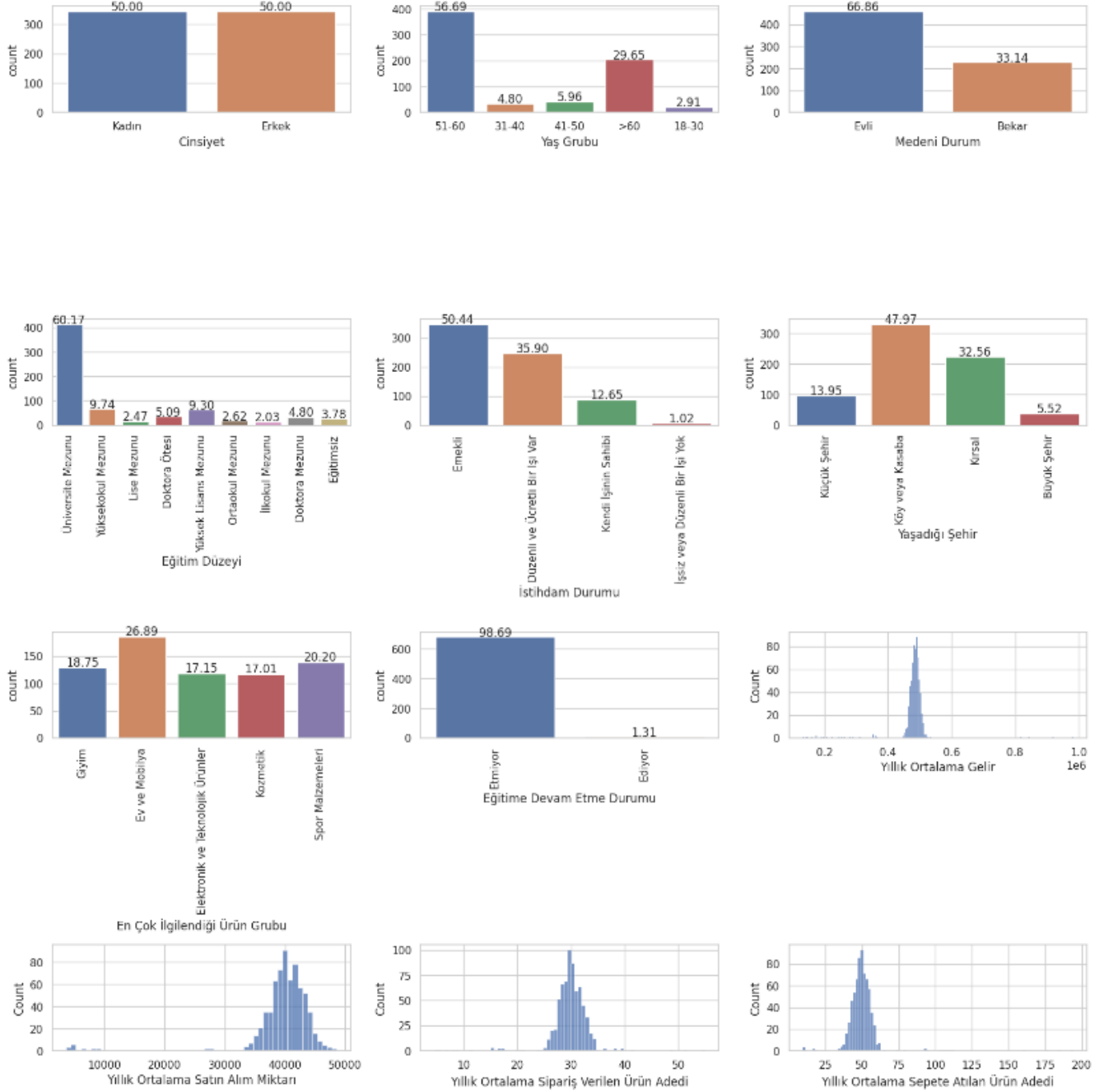
Bu gruptaki bireyler genellikle büyük şehirlerde yaşamakta, hem yüksek gelir hem de yüksek satın alma miktarına sahiptirler. Sipariş edilen ürün sayısı ile sepete eklenen ürün sayısı arasında doğrudan bir ilişki vardır ve bu miktarlar yüksek seviyelerdedir. Bu verilere dayanarak, ekonomik zorluk yaşamayan ve çeşitli ürün kategorilerine kolay erişimi olan bir kesimi yansıttığını düşündüğüm için bu grubu "Bolluk İçinde Yaşayanlar" olarak adlandırıyorum.



Şekil 21. Obek 5 İstatistikleri

Bu grubu oluşturan bireylerin çoğunluğu (%81.59) erkektir. Bu bireyler genellikle 18-40 yaş aralığındadır ve %56.23'ü bekârdır. Eğitim düzeyleri incelendiğinde, %75.64'ünün lise mezunu olduğu görülmektedir. İstihdam durumuna baktığımızda, %60.62'si düzenli ve maaşlı işlerde çalışmakta olup, genellikle küçük kasabalarda yaşamaktadırlar. Bu grubun en çok ilgilendiği ürün kategorileri spor malzemeleri, elektronik ve teknolojik ürünlerdir.

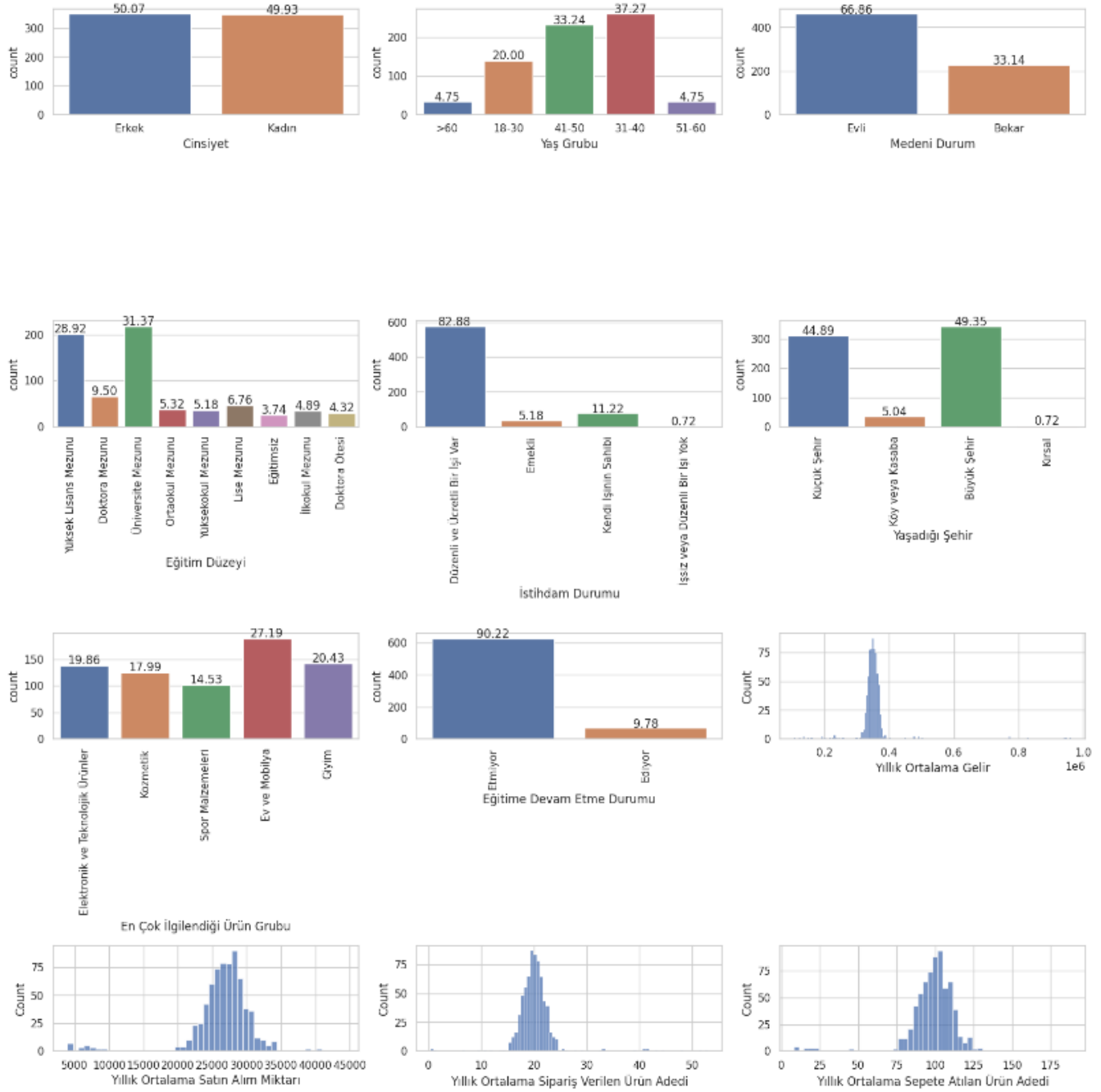
Yıllık ortalama gelirleri ve satın alma miktarları düşük seviyededir. Az sayıda ürün sipariş etmelerine rağmen, sepete daha fazla ürün eklemektedirler. Bu grubun en belirgin özelliği, genellikle mavi yakalı çalışanlar olarak bilinen çalışan erkek modelini temsil etmeleridir. Bu nedenle, bu grubun azimli ve kararlı doğasını yansıttığını düşündüğüm için onları "Mücadeleci Ruha Sahip Olanlar" olarak adlandırıyorum.



Şekil 22. Obek 6 İstatistikleri

Bu grubu oluşturan bireyler arasında kadınlar ve erkekler eşit şekilde temsil edilmektedir. Ayrıca, bu grubun bir parçası olan bireyler arasında 51 yaş ve üzerindeki kişiler bulunmaktadır ve grubun %66.86'sı evlidir. Üniversite eğitimi almış bireyler genellikle emekli durumdadır. Yıllık ortalama gelirleri normal seviyelerde olmasına rağmen, satın alma miktarları oldukça yüksektir. Hem sipariş edilen ürün sayısı hem de sepete eklenen ürün sayısı büyük miktardadır.

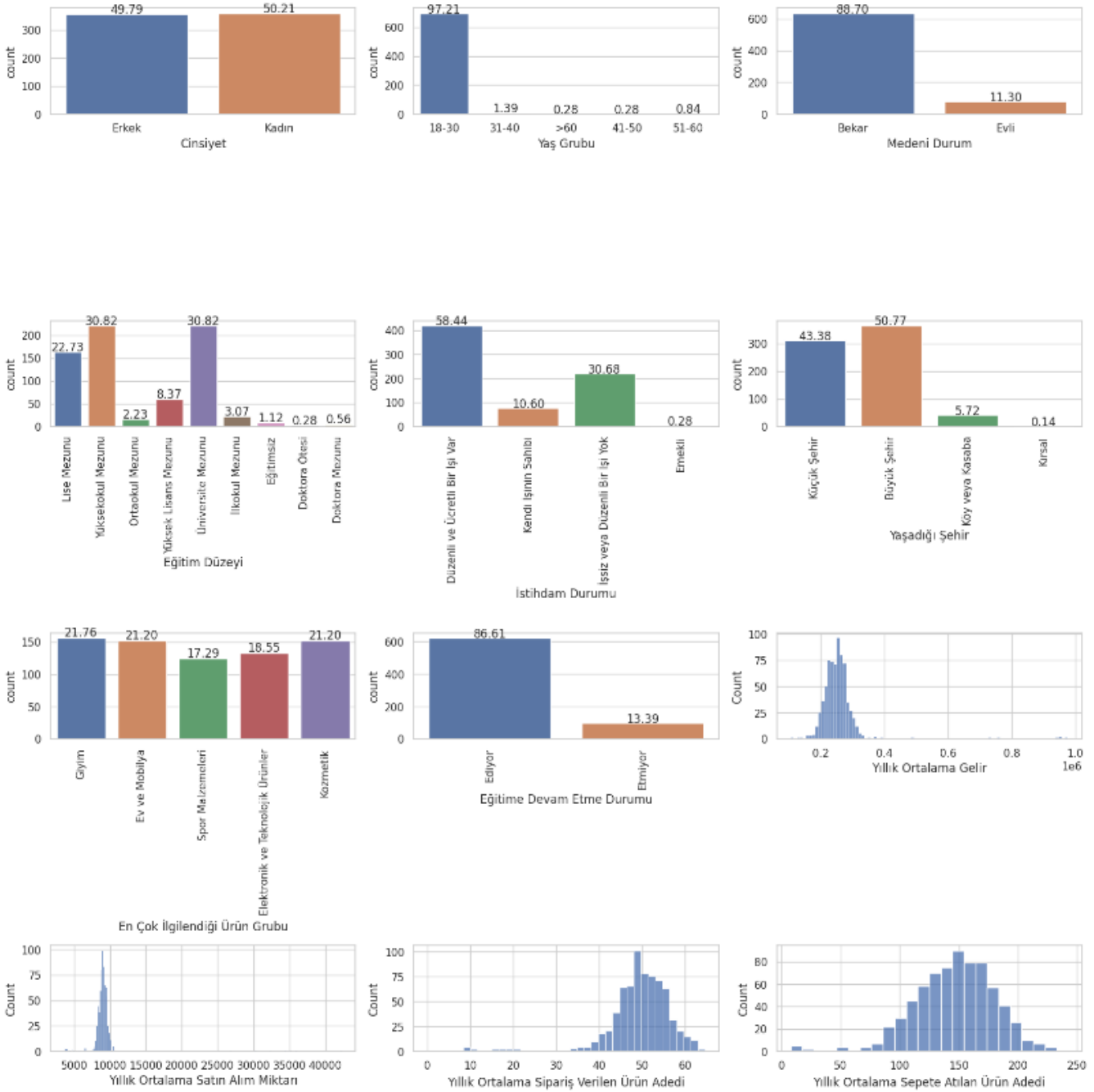
Bu grubun en belirgin özelliği, yaşlı yaş grubuna ve emeklilere odaklanmasıdır. Bu nedenle, bu gruba "Hayatın Tadını Çıkarıcılar" ismini vermeyi uygun bulduk.



Şekil 23. Obek 7 İstatistikleri

Kadın ve erkek bireylerden oluşan bu grup, cinsiyet açısından dengeli bir dağılım sergilemektedir. Ağırlıklı olarak 18-50 yaş aralığındaki bireylerden oluşan bu grup, evli kişilerle karakterize edilmektedir. Üniversite ve lisansüstü derecelere sahip bireylerden oluşan bu grupta, %82.88 oranında düzenli maaşlı işlerde çalışan bireyler bulunmaktadır. Bu kişiler, yoğun olarak büyük ve küçük şehirlerde yaşamaktadır ve ev ile mobilya ürünlerine yüksek ilgi göstermektedir.

Yıllık ortalama gelirleri genel ortalamaya uygun olsa da, dikkat çekici derecede yüksek satın alma miktarları öne çıkmaktadır. Sepete eklenen ürün sayısı fazla olmasına rağmen, sipariş edilen ürün sayısı daha düşüktür. Bu grubun en belirgin özelliği, eğitilmiş geçmişleri ve istikrarlı profesyonel kariyerleridir. Bu nedenle, bu grubu beyaz yakalı çalışanları temsil eden "Ofis Elitleri" olarak adlandırdım.



Şekil 24. Obek 8 İstatistikleri

Kadın ve erkek bireyler eşit şekilde temsil edilmektedir. Bu grubun yaş aralığı 18 ile 30 arasında olup, çoğunluğu bekârlardan oluşmaktadır. Eğitim seviyeleri lise, meslek okulu ve üniversite mezunlarını içermektedir. İstihdam durumuna baktığımızda, düzenli maaşlı işlerde çalışan bireylerin oranı yüksek (%58.44), ancak işsizlik oranı da (%30.68) dikkat çekicidir. Yıllık ortalama gelirleri ve satın alma miktarları düşük seviyededir. Ancak, satın alınan ürünlerin miktarı ve sepete eklenen ürünlerin sayısı yüksektir.

Bu grup ağırlıklı olarak genç nesli ve öğrencileri temsil etmektedir. Bu nedenle, bu grubu "Geleceğin Liderleri" olarak adlandırıyorum.

5 Deneyisel Calismalar

Bu bölümde, yedi farklı sınıflandırma modelini kullanarak bu modellerin sınıflandırma performanslarını değerlendireceğiz. Kullanılacak modeller şunlardır:

- * K-En Yakın Komşular (K-Nearest Neighbors),
- * Destek Vektör Makineleri (Support Vector Machine),
- * Lojistik Regresyon (Logistic Regression),
- * Naive Bayes,
- * Rastgele Orman (Random Forest),
- * Gradient Boosting,
- * Extreme Gradient Boosting (XGBoost).

Bu yöntemler **3.2 bölümünde** anlatılmıştır.

Her bir modelin performansı analiz edilecektir. Analiz sırasında accuracy, precision, recall, f1-score ve support değerleri dikkate alınacaktır.

5.1 Performans Metrikleri

Precision: Bir modelin pozitif olarak tahminlediği örnekler içindeki gerçekten pozitif olanların oranını gösterir. Yani, modelin “pozitif” dediği durumlar içerisinde ne kadar doğru olduğuna odaklanır. Böylece, yanlış alarm verme eğilimini ölçer.

$$\text{Precision} = \frac{TP}{TP + FP}$$

Denklem 8. Precision formulu

Recall: Recall (duyarlılık veya hatırlama oranı), gerçek pozitif örneklerin içinden modelin doğru bir şekilde pozitif olarak tespit edebildiklerinin oranını temsil eder. Bu metrik, modelin pozitif örnekleri ne ölçüde yakalayabildiğini, yani eksik tespit (kaçırma) oranının ne kadar düşük olduğunu gösterir.

$$\text{Recall} = \frac{TP}{TP + FN}$$

Denklem 9. Recall formulu

F1-Score: F1 skoru, precision ve recall’un harmonik ortalamasıdır. Bu skor, her iki metriğin de dengeli olarak yüksek olmasını vurgular ve dengesiz dağılımlı veri setlerinde, tek başına precision veya recall’a bakmaktan daha bütüncül bir değerlendirme sunar.

$$F1 = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

Denklem 9. F1-Score formulu

Support: Support, her bir sınıf için veri setindeki örnek sayısını ifade eder. Bu sayı, performans metriklerinin yorumlanmasında rehberlik eder çünkü az örnekli sınıflar için elde edilen sonuçlar istatistiksel olarak daha az güvenilir olabilir.

$$\text{Support} = \text{Sınıftaki örnek sayısı}$$

Denklem 10. Support formulu

Accuracy: Accuracy, tüm örnekler içindeki doğru sınıflandırılan örneklerin oranıdır. Yani modelin genel olarak ne kadar doğru sonuç ürettiğini basitçe ölçer. Ancak sınıf dağılımları dengesiz olduğunda, tek başına accuracy yanıltıcı olabilir.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

Denklem 11. Accuracy formulu

Confusion matrix: Confusion matrix, bir sınıflandırma modelinin gerçek sınıflarla tahmin edilen sınıflar arasındaki ilişkiyi gösteren basit bir tablodur. Satırlar gerçek etiketleri, sütunlar modelin tahmin ettiği etiketleri temsil eder. Bu sayede modelin hangi durumlarda doğru tahmin yaptığı, hangi durumlarda hangi sınıfları karıştırdığı net bir şekilde görülebilir.

	Tahmin: Pozitif	Tahmin: Negatif
Gerçek: Pozitif	TP	FN
Gerçek: Negatif	FP	TN

Denklem 12. Confusion matrix yapısı

5.2 Modellerin Sonuc Uzerine Etkisi

5.2.1 Logistic Regression

	precision	recall	f1-score	support
0	0.96	0.90	0.92	143
1	0.95	0.98	0.96	143
2	0.94	0.99	0.96	144
3	0.93	0.97	0.95	144
4	0.96	0.95	0.95	144
5	0.94	0.95	0.94	143
6	0.95	0.96	0.95	144
7	0.99	0.92	0.95	143
accuracy			0.95	1148
macro avg	0.95	0.95	0.95	1148
weighted avg	0.95	0.95	0.95	1148

```
[[128 3 3 3 1 2 3 0]
 [ 0 140 1 1 0 1 0 0]
 [ 1 1 142 0 0 0 0 0]
 [ 0 1 0 139 2 2 0 0]
 [ 3 0 0 1 137 2 1 0]
 [ 1 0 0 4 0 136 2 0]
 [ 1 1 1 0 2 0 138 1]
 [ 0 2 4 1 1 2 2 131]]
LR accuracy: 95.03%
```

Tablo 1. Logistic Regressionun Sonuc Tablosu

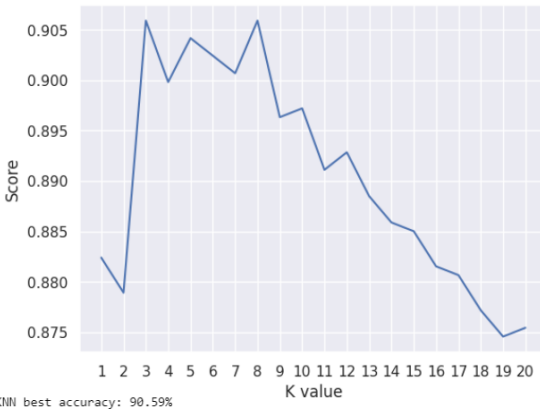
5.2.4 Naive Bayes

	precision	recall	f1-score	support
0	0.92	0.91	0.91	143
1	0.95	0.98	0.97	143
2	0.94	0.99	0.96	144
3	0.96	0.96	0.96	144
4	0.96	0.91	0.93	144
5	0.93	0.98	0.95	143
6	0.96	0.96	0.96	144
7	0.99	0.92	0.95	143
accuracy			0.95	1148
macro avg	0.95	0.95	0.95	1148
weighted avg	0.95	0.95	0.95	1148

```
[[130 3 3 4 1 1 1 0]
 [ 0 140 1 1 0 1 0 0]
 [ 1 1 142 0 0 0 0 0]
 [ 0 1 0 138 2 3 0 0]
 [ 9 0 0 0 131 3 1 0]
 [ 1 0 0 0 0 140 2 0]
 [ 1 1 1 0 2 0 138 1]
 [ 0 1 4 1 1 3 2 131]]
Gaussian Naive Bayes accuracy: 94.95%
```

Tablo 4. Naive Bayes Sonuc Tablosu

5.2.2 K-Nearest Neighbour (KNN)



Tablo 2. KNN'nin Sonuc Grafigi

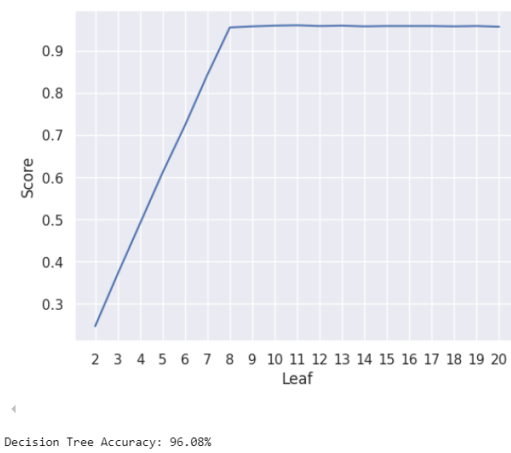
5.2.3 Support Vector Machine

	precision	recall	f1-score	support
0	0.98	0.91	0.94	143
1	0.95	0.98	0.97	143
2	0.94	0.99	0.96	144
3	0.95	0.97	0.96	144
4	0.96	0.97	0.97	144
5	0.96	0.98	0.97	143
6	0.96	0.96	0.96	144
7	0.99	0.93	0.96	143
accuracy			0.96	1148
macro avg	0.96	0.96	0.96	1148
weighted avg	0.96	0.96	0.96	1148

```
[[130 3 3 4 1 1 1 0]
 [ 0 140 1 1 0 1 0 0]
 [ 1 1 142 0 0 0 0 0]
 [ 0 1 0 140 2 1 0 0]
 [ 0 0 0 0 140 3 1 0]
 [ 1 0 0 0 0 140 2 0]
 [ 1 1 1 0 2 0 138 1]
 [ 0 1 4 2 1 0 2 133]]
SVC accuracy: 96.08%
```

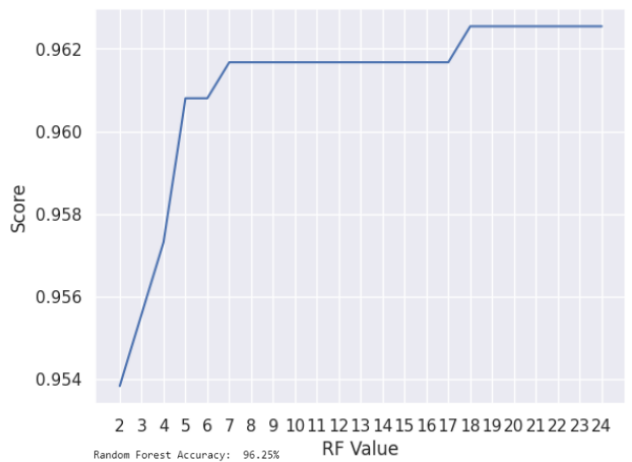
Tablo 3. Support Vector Machine Sonuc Tablosu

5.2.5 Desicion Tree



Tablo 5. Desicion Tree Sonuc Grafigi

5.2.6 Random Forest



Tablo 6. Random Forest Sonuc Grafigi

5.2.7 XGBoost

	precision	recall	f1-score	support
0	0.98	0.91	0.94	143
1	0.95	0.98	0.96	143
2	0.91	0.99	0.95	144
3	0.97	0.97	0.97	144
4	0.96	0.98	0.97	144
5	0.95	0.97	0.96	143
6	0.96	0.94	0.95	144
7	0.99	0.92	0.95	143
accuracy			0.96	1148
macro avg	0.96	0.96	0.96	1148
weighted avg	0.96	0.96	0.96	1148

```
[[130 3 3 4 1 1 1 0]
 [ 0 140 1 1 0 1 0 0]
 [ 1 1 142 0 0 0 0 0]
 [ 0 2 0 139 2 1 0 0]
 [ 0 0 0 0 141 3 0 0]
 [ 1 1 0 0 0 138 2 1]
 [ 1 1 3 0 2 0 136 1]
 [ 0 0 7 0 1 1 2 132]]
Gradient Boosting accuracy: 95.64%
```

Tablo 7. XGBoost Sonuc Grafigi

5.3 Modellerin Karsilastirilmesi

	Model	Accuracy
5	Random Forest	96.254355
2	SVM	96.080139
4	Decision Tree	96.080139
6	Gradient Boost	95.644599
0	Logistic Regression	95.034843
3	Gaussian NB	94.947735
1	K Neighbors	90.592334

Tablo 8. Modellerin Karsilastirilmesi

Random Forest modelinin sınıflandırma doğruluğu açısından diğer modellerden daha iyi performans gösterdiği açıktır. Sınıflandırma sonuçlarını daha da iyileştirmek için, veri seti üzerinde bir özellik seçimi tekniği uygulayacak ve Rastgele Orman modeli için hiper-parametreleri belirleyeceğiz.

5.4 Rastgele Orman İçin Hiperparametre Ayarı

Sınıflandırma sürecinde, Rastgele Orman algoritması içinde varsayılan parametreler kullanılmıştır, ancak maksimum yaprak düğümleri hariç tutulmuştur. Şu anda, parametreleri ince ayar yapmak için grid search (ızgara arama) yöntemi kullanarak parametre seçimini optimize etmeyi amaçlıyoruz.

Grid Search: Grid Search, bir makine öğrenimi modelinin en iyi performansı verecek hiperparametre değerlerini belirlemek için

sistematiik olarak tüm kombinasyonları tarayan bir arama yöntemidir. Modelin ayarlanabilecek parametreleri için önceden belirlenmiş bir değer kümesi oluşturulur ve bu değerlerin her olası kombinasyonu denenir. Her bir kombinasyon için model eğitilip değerlendirilir, sonuçlar kaydedilir. Daha sonra performans metriklerine göre en başarılı kombinasyonlar seçilir. Bu sayede, modelin eldeki verisetindeki başarısını arttıracak en uygun hiperparametreler, deneme-yanılma süreciyle ancak metodik ve sistematiik bir yaklaşımla tespit edilebilir.

```
n_estimators = [int(x) for x in np.linspace(start = 10, stop = 80, num = 10)]
max_features = ['sqrt']
max_depth = [20]
min_samples_split = [2, 5]
min_samples_leaf = [1, 2]
bootstrap = [True, False]

param_grid = {'n_estimators': n_estimators,
              'max_features': max_features,
              'max_depth': max_depth,
              'min_samples_split': min_samples_split,
              'min_samples_leaf': min_samples_leaf,
              'bootstrap': bootstrap}

print(param_grid)

['n_estimators': [10, 17, 25, 33, 41, 48, 56, 64, 72, 80], 'max_features': ['sqrt'], 'max_depth': [20], 'min_samples_split': [2, 5], 'min_samples_leaf': [1, 2], 'bootstrap': [True, False]]

def grid_search_cv(model):
    rf_grid = GridSearchCV(estimator = model, param_grid = param_grid, cv = 5, verbose=0, n_jobs = 4)
    rf_grid.fit(X_train, y_train)
    return rf_grid.best_params_
best_params = grid_search_cv(RFClassifier)
print('Best parameters are: ', best_params)

Best parameters are: {'bootstrap': True, 'max_depth': 4, 'max_features': 'sqrt', 'min_samples_leaf': 1, 'min_samples_split': 2, 'n_estimators': 17}

RFClassifier=RandomForestClassifier(bootstrap=True, max_depth=4, max_features='sqrt', min_samples_leaf=1, min_samples_split=2, n_estimators=17)
RFClassifier.fit(X_train, y_train)
RFClassifier.predict(X_test)
result = accuracy_score(y_test, result)
print('Random Forest Classification Accuracy: {:.2f}%'.format(result*100))

Random Forest Classification Accuracy: 96.17%
```

Şekil 25. Hiperparametre Ayarı

n_estimators: Ormanda kaç adet karar ağacı olacağını belirler. Daha fazla ağaç genellikle daha istikrarlı bir model, ancak daha uzun eğitim süresi anlamına gelir.

max_features: Her bir ağacı oluştururken her düğümde seçilecek özellik sayısını belirler. Örneğin "sqrt" genellikle karesel kök oranda özelliği seçeceği anlamına gelir.

max_depth: Ağaçların dallanabileceği maksimum derinliktir. Bu değer, ağaçların aşırı öğrenmesini (overfitting) engellemek için sınırlanabilir.

min_samples_split: Bir düğümü bölmek için gereken minimum örnek sayısını belirler. Daha büyük bir sayı, ağaçların daha genel, daha az dallanmış olmasını sağlar.

min_samples_leaf: Bir yaprak düğümde bulunması gereken minimum örnek sayısıdır. Bu değer de ağaçların fazla dallanmasını önler.

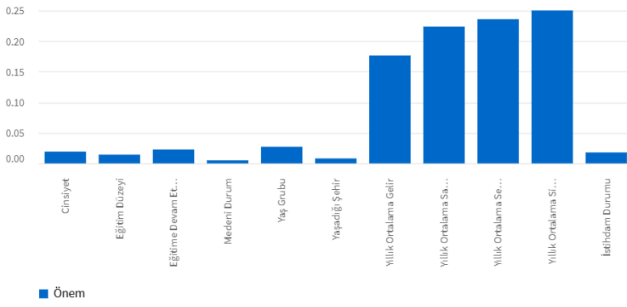
bootstrap: Her ağacı oluştururken, verisetinden örneklerin tekrar seçilip seçilmeyeceğini (bootstrap örnekleme) belirler. True değerinde örnekler tekrar seçilerek alınırken, False olduğunda bu yapılmaz.

Hiperparametreleri modele eklediğimizde, sınıflandırma doğruluğunun azaldığını görebiliriz. Modele ek parametreler dahil edildiğinde, bu parametrelerin getirdiği artan karmaşıklık nedeniyle sınıflandırma doğruluğu düşebilir. Bu durum, modelin aşırı öğrenmesi (overfitting) ile açıklanabilir. Aşırı öğrenme, modelin eğitim verisine gereğinden fazla uyum sağlaması, dolayısıyla verideki gürültüyü de öğrenerek yeni ve görülmemiş verilere genelleme yapma kabiliyetinin azalması anlamına gelir.

5.5 Özellik Secimi

Modelin hangi özelliklere (feature) daha fazla önem verdiğini göstermek için "feature importance" değerlerini ekrana yansıtan bir bölüm eklenmiştir. Bu sayede kullanıcının girdiği verilerin model tahminlerine hangi oranda katkıda bulunduğu genel olarak görülebilir.

```
feature_importances = rf_model.feature_importances_  
fi_df = pd.DataFrame({'Özellik': feature_names, 'Önem': feature_importances})  
fi_df = fi_df.sort_values('Önem', ascending=False)
```



Şekil 26. Feature Importances Tablosu

6 Sonuc

Bu çalışma, müşteri segmentasyonu ile karar destek sistemlerinin nasıl geliştirilebileceğini kapsamlı bir şekilde ele almış ve bu bağlamda veri analitiği, makine öğrenimi ve kullanıcı odaklı tasarım prensiplerini bir araya getirmiştir. Random foresti kullanarak girilen kullanıcı bilgilerine göre öbek tahmini yapılmıştır. Her bir öbek için kişiselleştirilmiş reklam stratejileri ve ürün önerileri geliştirilmiş, böylece pazarlama yaklaşımlarının etkinliği artırılmıştır.

Ayrıca, çalışmada kullanılan veri seti analiz edilerek eksiksizlik ve tutarlılık sağlanmış, sürekli ve kategorik değişkenlerin dağılımları detaylı

olarak incelenmiştir. Veri görselleştirme ve kullanıcı etkileşimi için Streamlit gibi modern araçlar kullanılmış, bu sayede kullanıcı dostu bir arayüz sunulmuştur. Bunun yanı sıra, uygulamanın CSS ile desteklenmiş olması, görselliği ve profesyonel bir görünümü artırmıştır.

Uygulamanın dinamik yapısı, kullanıcıdan gelen yeni verilerin modele entegre edilmesini ve modelin zaman içinde yeniden eğitilmesini mümkün kılarak sürekli bir öğrenme mekanizması oluşturmuştur. Bu durum, pazarlama stratejilerinin güncel verilere dayanarak daha etkili ve hedefe yönelik hale gelmesini sağlamıştır. Çalışmada ayrıca, farklı makine öğrenimi algoritmaları karşılaştırılmış ve Random Forest modelinin doğruluk açısından üstünlüğü vurgulanmıştır.

Sonuç olarak, bu proje, müşteri segmentasyonu ve karar destek sistemleri alanında yenilikçi bir yaklaşım sunmuş, işletmelerin veriye dayalı karar alma süreçlerini optimize etme potansiyelini ortaya koymuştur. Önerilen yöntemler, sadece pazarlama stratejileri için değil, aynı zamanda müşteri memnuniyetini artırmaya yönelik uygulamalarda da kullanılabilir, bu da çalışmanın geniş bir etki alanına sahip olduğunu göstermektedir.

Kaynaklar

- [1] Budak, H., & Gümüştas, E. (2022). Kişiselleştirilmiş ürün öneri sistemi için kullanıcı bazlı işbirlikçi filtreleme ve kümeleme kullanan hibrit bir yaklaşım. İstanbul Ticaret Üniversitesi Sosyal Bilimler Dergisi, 21(43), 253-268. doi: 10.46928/iticusbe.1055162
- [2] Sinap, V. (2024). Perakende Sektöründe Makine Öğrenmesi Algoritmalarının Karşılaştırmalı Performans Analizi: Black Friday Satış Tahminlemesi. Selçuk Üniversitesi Sosyal Bilimler Meslek Yüksekokulu Dergisi, 27 (1), 65-90
- [3] Ergün, O. (2023). *Makine öğrenmesi algoritmaları ile müşteri segmentasyonu ve hepsiburada E-ticaret platformu üzerine bir uygulama* (Order No. 30721506). Available from ProQuest Dissertations & Theses Global. (2890697795). Retrieved from <https://www.proquest.com/dissertations-theses/makine-ogrenmesi-algoritmaları-ile-müşteri/docview/2890697795/se-2>
- [4] *Yapay Sinir Ağları*. İstanbul: Papatya yayıncılık, 2. baskı, .
- [5] Kutlugün ve ark. (2017). *Yapay Sinir Ağları ve K-En Yakın Komşu Algoritmalarının Birlikte Çalışma Tekniği (Ensemble) ile Metin Türü Tanıma*. İstanbul: XXII. Türkiye’de Internet Konferansı (inet’tr17), Bahçeşehir Üniversitesi,.