

T.C.

BİTLİS EREN ÜNİVERSİTESİ
MÜHENDİSLİK-MİMARLIK FAKÜLTESİ



Doğal Dil İşleme Final Proje Raporu

Öğrenci No: 21080410021

Ad, Soyadı: EYYÜP KARDEŞ

KONU: Amazon ürün veri setini kullanarak ürünleri kategorilere ayırabilen bir model oluşturma

VERİ SETİ HAKKINDA BİLGİ:

Kaynak: Amazon.in'den çekilmiştir.

Boyut: 20.000'den fazla satır ve 9 sütun.

Sütunlar:

name: Ayakkabının adı.

main_category: Ayakkabının ana kategorisi (çoğunlukla "Kadın Ayakkabıları").

sub_category: Ayakkabının alt kategorisi (örneğin "Ayakkabılar", "Sandaletler").

image: Ayakkabının resim URL'si.

link: Ayakkabının Amazon.in'deki ürün sayfasının URL'si.

ratings: Ayakkabının ortalama puanı (5 üzerinden).

no_of_ratings: Ayakkabıya verilen toplam puan sayısı.

discount_price: Ayakkabının indirimli fiyatı (₹).

actual_price: Ayakkabının gerçek fiyatı (₹).

Veri Seti Hakkında:

Veri seti, çeşitli kadın ayakkabı modellerini içermektedir.

Ayakkabıların çoğu, çeşitli alt kategorilerde ("Ayakkabılar", "Sandaletler", vb.) sınıflandırılmıştır.

Her ayakkabı için bir resim URL'si ve ürün sayfası URL'si mevcuttur.

Puanlar ve puan sayıları, ayakkabıların popüleritesi ve müşteri memnuniyeti hakkında bilgi sağlar.

İndirimli ve gerçek fiyatlar, fiyat analizi ve karşılaştırma yapma imkânı sunar.

Kullanım Alanları:

Marka Tahmini: Ayakkabı adını kullanarak markayı tahmin etmek için bir model eğitilebilir.

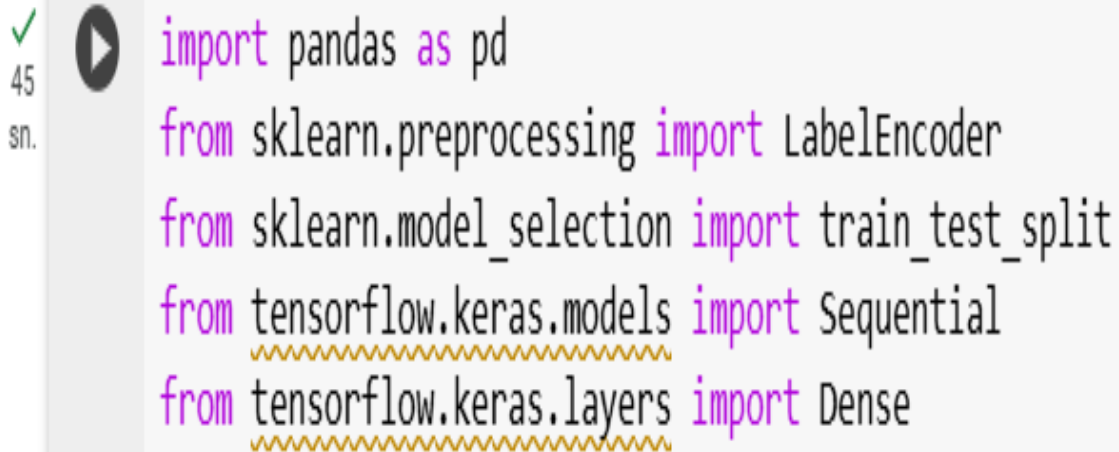
Kategori Tahmini: Ayakkabı özelliklerini kullanarak kategoriyi tahmin etmek için bir model eğitilebilir.

Fiyat Tahmini: Ayakkabı özelliklerini kullanarak fiyatı tahmin etmek için bir model eğitilebilir.

Puan Tahmini: Ayakkabı özelliklerini ve yorumlarını kullanarak puanı tahmin etmek için bir model eğitilebilir.

Öneri Sistemi: Müşterilerin beğenebileceği ayakkabıları önermek için bir öneri sistemi oluşturulabilir.

MODEL OLUŞTURMA



```
import pandas as pd
from sklearn.preprocessing import LabelEncoder
from sklearn.model_selection import train_test_split
from tensorflow.keras.models import Sequential
from tensorflow.keras.layers import Dense
```

Bu bölümde, kodun çalışması için gerekli kütüphaneler içe aktarılıyor.

pandas:

Verileri okumak, işlemek ve analiz etmek için kullanılır.

sklearn.preprocessing:

Verileri modele uygun hale getirmek için kullanılır. LabelEncoder, kategorik verileri (örneğin marka isimleri) sayılara dönüştürür.

sklearn.model_selection:

train_test_split fonksiyonu ile veri setini eğitim ve test kümelerine ayırır.

tensorflow.keras:

Yapay sinir ağı modelini oluşturmak ve eğitmek için kullanılır. Sequential, modeli katman katman oluşturmayı sağlar. Dense ise standart bir yapay sinir ağı katmanıdır

Veri Yükleme ve Hazırlama

```
# Dosya adını al
file_name = list(uploaded.keys())[0]

# "Shoes.csv" dosyasını oku
df = pd.read_csv(file_name)

# İlk 5 satırı göster
print(df.head().to_markdown(index=False, numalign="left", stralign="left"))

# Sütunları ve türlerini göster
print(df.info())

# Her satırdaki ilk kelimeyi çıkar
df['brand'] = df['name'].str.split().str[0]

# One-hot kodlama
X = df['brand']
y = pd.get_dummies(df['main_category'])

# LabelEncoder oluştur
le = LabelEncoder()

# X'i dönüştür
X = le.fit_transform(X)

# Verileri eğitim ve test kümelerine ayır
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)
```

Bu bölümde, ayakkabı veri seti yükleniyor ve modele uygun hale getiriliyor.

Google Colab kullanılıyorsa, dosya yüklenir.

`pd.read_csv(file_name)` ile veri seti okunur ve bir pandas DataFrame'ine yüklenir.

`df.head()` ve `df.info()` ile veri seti hakkında bilgi edinilir.

`df['brand'] = df['name'].str.split().str[0]` ile her ayakkabı adının ilk kelimesi marka olarak kabul edilir ve yeni bir "brand" sütununa eklenir.

`pd.get_dummies(df['main_category'])` ile "main_category" sütunu one-hot encoding yöntemiyle sayısal hale getirilir.

LabelEncoder kullanılarak "brand" sütunundaki markalar sayısal olarak etiketlenir.

`train_test_split` ile veriler eğitim ve test kümelerine ayrılır.

Model Oluşturma ve Eğitim

Bu bölümde, yapay sinir ağı modeli oluşturuluyor ve

```
# YSA modelini oluştur
model = Sequential()
model.add(Dense(128, activation='relu', input_shape=(1,)))
model.add(Dense(64, activation='relu'))
model.add(Dense(y.shape[1], activation='softmax'))

# Modeli derle
model.compile(optimizer='adam', loss='categorical_crossentropy', metrics=['accuracy'])

# Modeli eğit
X_train = X_train.reshape(-1, 1)
X_test = X_test.reshape(-1, 1)
model.fit(X_train, y_train, epochs=10, batch_size=32)
```

eğitiliyor.

Sequential() ile bir yapay sinir ağı modeli oluşturulur.

Üç katmanlı bir model tanımlanır. İlk iki katman ReLU aktivasyon fonksiyonunu kullanır.

Son katman, softmax aktivasyon fonksiyonunu kullanarak çıktıları olasılıklara dönüştürür.

model.compile() ile model derlenir. "adam" optimizer'ı, "categorical_crossentropy" kayıp fonksiyonu ve "accuracy" metriği kullanılır.

model.fit() ile model eğitilir.

Model Değerlendirme ve Karışıklık Matrisi

```
# Ta # Test verilerindeki gerçek sınıfları dönüştür
y_pr y_test_classes = pd.DataFrame(y_test).idxmax(axis=1)

# Do # Karışıklık matrisini hesapla ve sınıf etiketlerini al
loss cm = confusion_matrix(y_test_classes, y_pred_classes, labels=y.columns)
prin

# Ka # Karışıklık matrisini çiz
from plt.figure(figsize=(10, 7))
imp sns.heatmap(cm, annot=True, fmt='d', cmap='Blues', xticklabels=y.columns, yticklabels=y.columns)
imp plt.xlabel('Tahmin Edilen Sınıf')

# Ta plt.ylabel('Gerçek Sınıf')
y_pr plt.title('Karışıklık Matrisi')
y_pr plt.show()

# Te
y_te
```

Dosyaları Seç Shoes.csv

• Shoes.csv(text/csv) - 560965 bytes, last modified: 12.12.2024 - 100% done

Bu bölümde, eğitilen modelin performansı değerlendirilir ve bir karışıklık matrisi çizilir.

`model.predict()` ile test verileri kullanılarak tahminler yapılır.

`model.evaluate()` ile modelin doğruluk skoru hesaplanır.

`confusion_matrix` fonksiyonu ile karışıklık matrisi oluşturulur.

seaborn ve matplotlib kütüphaneleri kullanılarak karışıklık matrisi görselleştirilir.

Bu kod parçasında, Amazon'dan alınan bir ayakkabı veri seti kullanılarak bir ayakkabı ana kategori tahmin modeli oluşturuluyor. Amaç, verilen bir ayakkabı markasına göre, ayakkabının ait olduğu ana kategoriyi (örneğin, kadın ayakkabıları, erkek ayakkabıları, çocuk ayakkabıları) tahmin edebilen bir yapay sinir ağı modeli eğitmek.

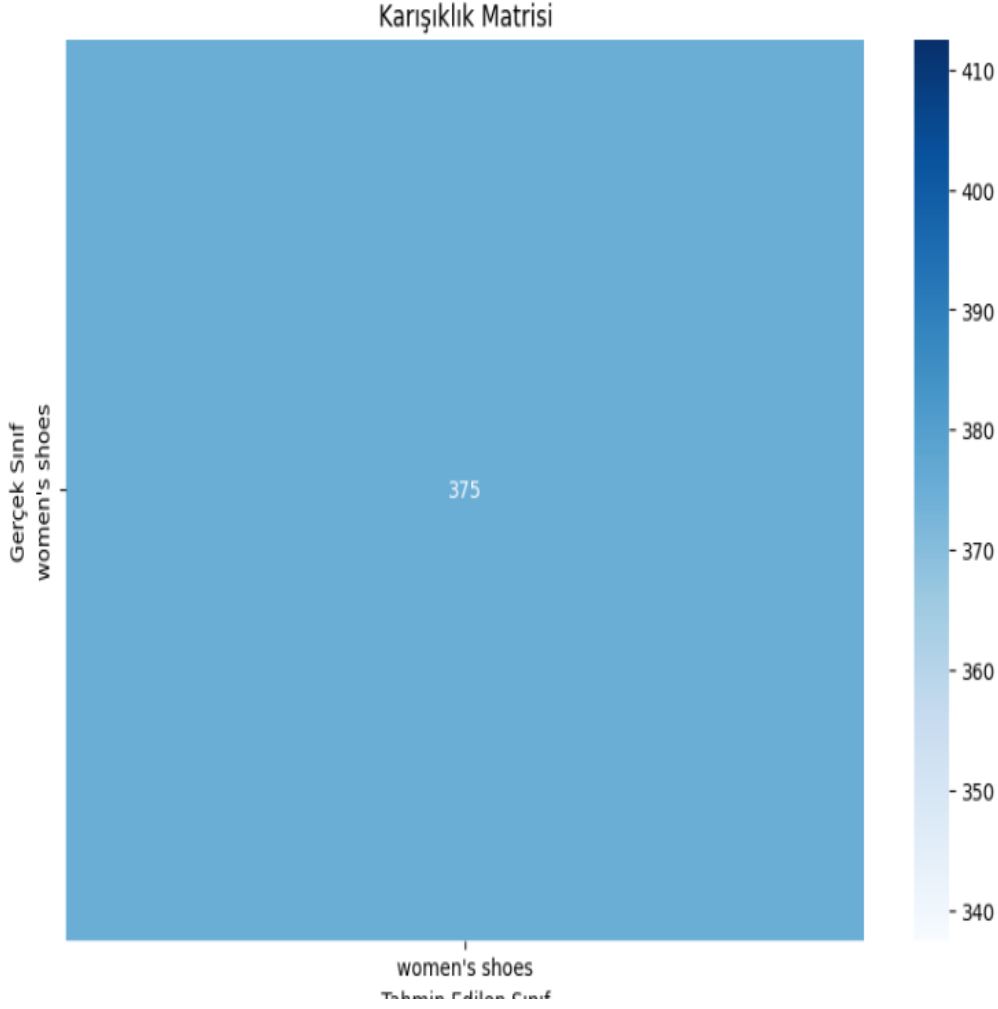
Kod, temel olarak şu adımları izliyor:

Veri Seti Hazırlama: Ayakkabı veri seti okunur ve gerekli ön işleme adımları uygulanır. Marka bilgisi ayrıştırılır ve hem marka hem de tahmin edilecek ana kategori bilgisi yapay sinir ağlarının anlayabileceği sayısal formata dönüştürülür.

Model Oluşturma: Yapay sinir ağı modeli tensorflow.keras kütüphanesi kullanılarak oluşturulur. Modelin katmanları ve nöron sayıları belirlenir, aktivasyon fonksiyonları seçilir.

Model Eğitimi: Hazırlanan veri setinin bir kısmı kullanılarak model eğitilir. Bu eğitim sürecinde, model marka ve ana kategori arasındaki ilişkiyi öğrenir.

Model Değerlendirmesi: Eğitilen model, veri setinin geri kalan kısmı üzerinde test edilir ve doğruluk oranı hesaplanır. Kısacası, bu kod ayakkabı markasını girdi olarak alıp, ayakkabının ait olduğu ana kategoriyi tahmin eden bir model oluşturmayı ve eğitmeyi hedefliyor.



Bu karışıklık matrisi, modelinizin "women's shoes" (kadın ayakkabıları) kategorisini tahmin etmedeki başarısını gösteriyor. Matrisin sadece bir hücresi olması, veri setinizde tek bir sınıfın (kadın ayakkabıları) bulunduğunu gösteriyor.

Bu karışıklık matrisi, modelinizin "women's shoes" (kadın ayakkabıları) kategorisini tahmin etmedeki başarısını gösteriyor. Matrisin sadece bir hücresi olması, veri setinizde tek bir sınıfın (kadın ayakkabıları) bulunduğunu gösteriyor.

Karışıklık Matrisinin Anlamı:

Satırlar: Gerçek sınıfları temsil eder. Bu durumda, tüm ayakkabıların gerçekte "women's shoes" kategorisine ait olduğunu biliyoruz.

Sütunlar: Modelin tahmin ettiği sınıfları temsil eder.

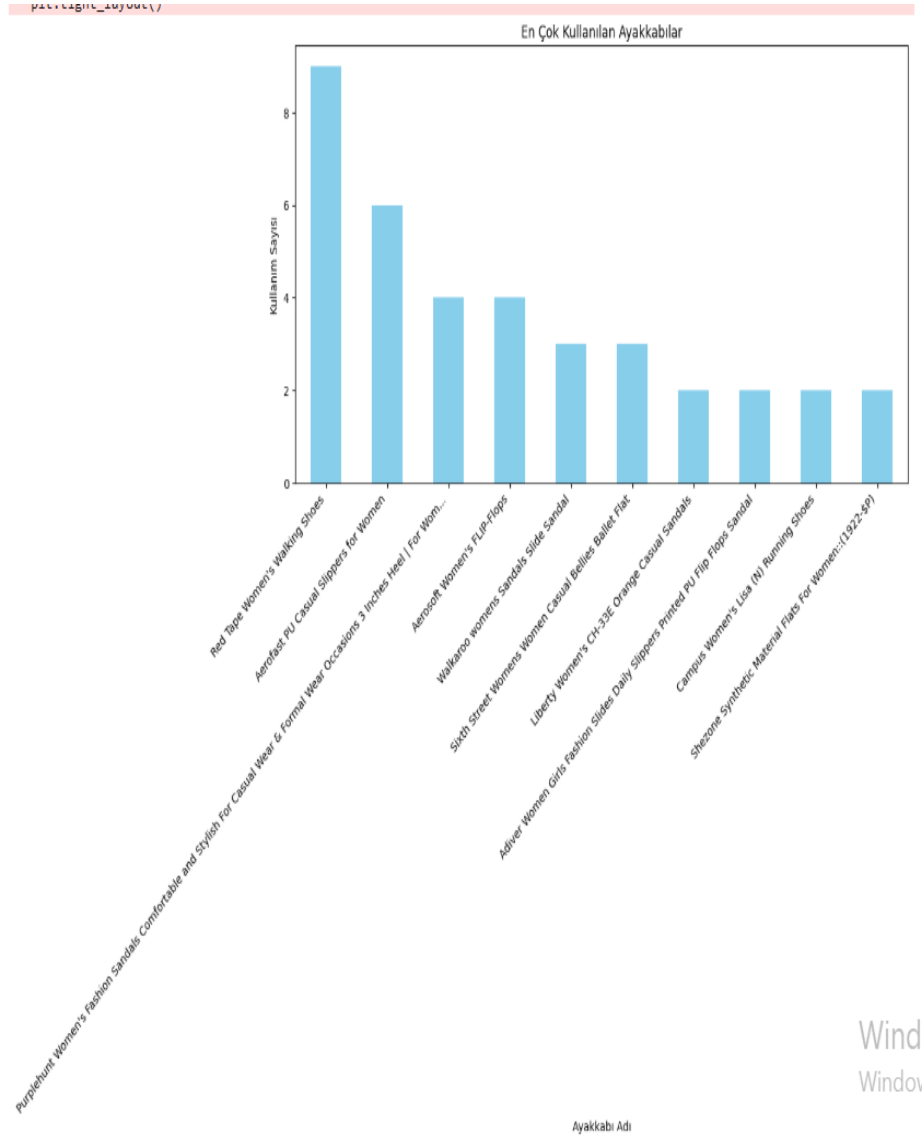
Hücre Değerleri: Her hücre, gerçek sınıf ile tahmin edilen sınıfın kombinasyonuna karşılık gelen ayakkabı sayısını gösterir.

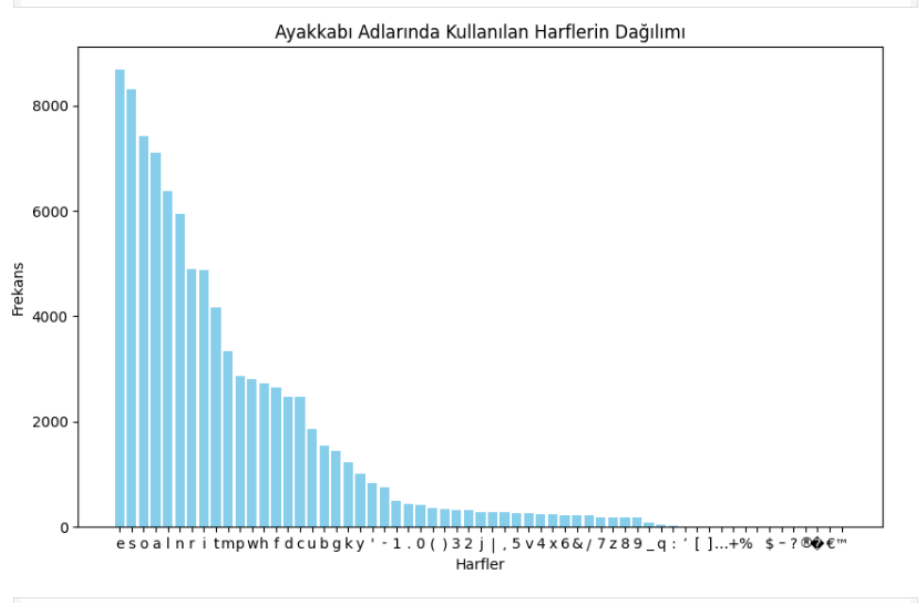
Matrisin Sonuçları:

375 ayakkabı doğru bir şekilde "women's shoes" olarak sınıflandırılmıştır (Doğru Pozitif).

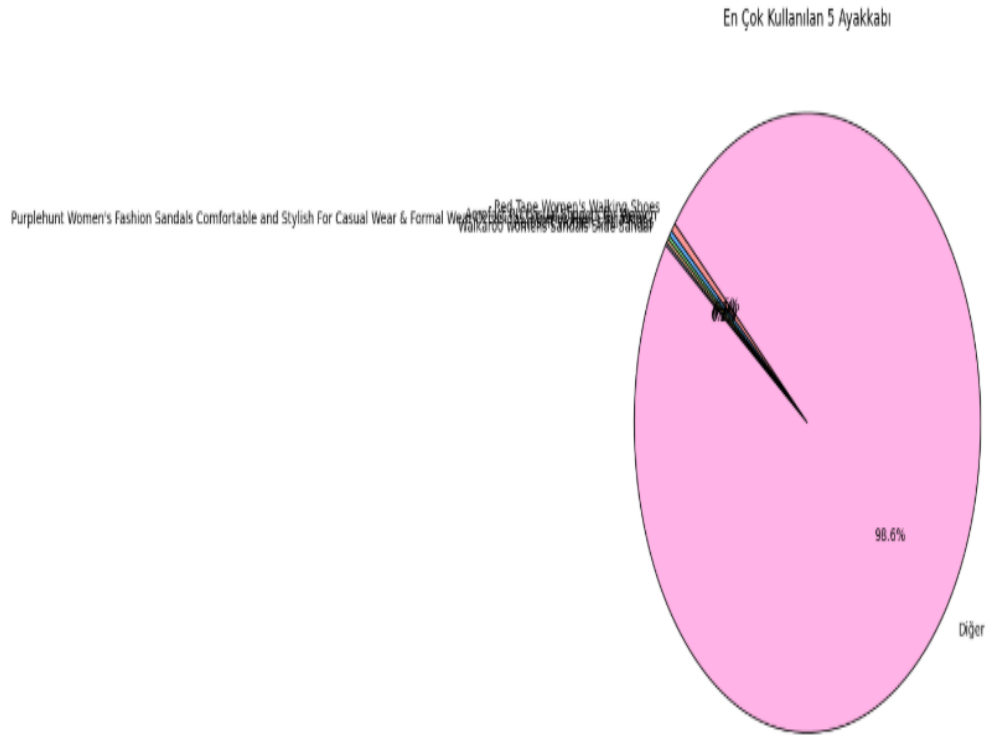
Başka bir deyişle, modeliniz tüm kadın ayakkabılarını doğru bir şekilde tanımlamıştır.

Bu kod, "Shoes.csv" adlı bir dosyadan ayakkabı verilerini alıp analiz ediyor. Analiz sonucunda hangi ayakkabının ne kadar kullanıldığını buluyor ve bunu hem liste halinde yazdırıyor hem de "sorted_shoes.csv" adlı yeni bir dosyaya kaydediyor. Son olarak, en çok kullanılan 10 ayakkabıyı gösteren bir bar grafiği çiziyor. Böylece hangi ayakkabıların daha popüler olduğunu kolayca görebiliyoruz.



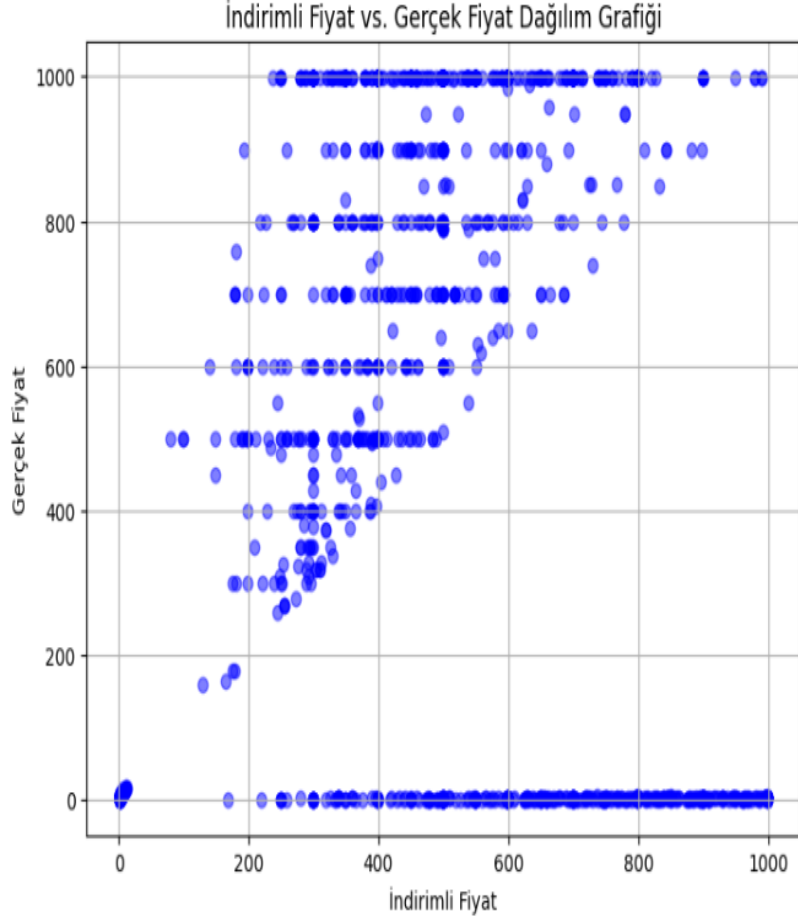


Bu grafik, ayakkabı adlarında hangi harflerin ne kadar sık kullanıldığını gösteriyor. "e" harfi en çok kullanılan harf iken, "q", "[" gibi bazı harfler neredeyse hiç kullanılmamış. Grafik sayesinde ayakkabı isimlerinde hangi harflerin daha popüler olduğunu görebiliyoruz.



Bu pasta grafiđi, en ok kullanılan 5 ayakkabının kullanım oranlarını gsteriyor. Grldđ gibi, "Purplehunt Women's Fashion Sandals Comfortable and Stylish For Casual Wear & Formal Wear" isimli ayakkabı ezici bir stnlkle en ok kullanılan ayakkabı. Diđer 4 ayakkabının kullanım oranları ise olduka dřk ve neredeyse fark edilemeyecek kadar kk dilimlerle temsil edilmiř. Pasta grafiđinin byk ođunluđunu kaplayan pembe alan "Diđer" kategorisini temsil ediyor, yani bu 5 ayakkabı dıřında kalan tm ayakkabıların toplam kullanım oranı.

Kısacası, bu grafik belirli bir ayakkabı modelinin ok popler olduđunu ve diđerlerinin kullanım oranlarının ihmal edilebilir dzeyde olduđunu gsteriyor.



Bu grafik, ayakkabıların indirimli fiyatları ile gerçek fiyatları arasındaki ilişkiyi gösteren bir **dağılım grafiği**. Her nokta bir ayakkabıyı temsil ediyor. Noktanın yatay konumu indirimli fiyatını, dikey konumu ise gerçek fiyatını gösteriyor.

Grafiğe bakarak şunları söyleyebiliriz:

- **Çoğu ayakkabının indirimli fiyatı, gerçek fiyatından düşük:** Grafikteki noktaların çoğu, grafiğin sol alt köşesinden sağ üst köşesine doğru uzanan hayali bir çizginin altında yer alıyor. Bu, indirimli fiyatın genellikle gerçektan daha düşük olduğunu gösteriyor.
- **Bazı ayakkabılar indirimsiz satılıyor:** Yatay eksene yakın olan noktalar, indirimli ve gerçek fiyatının

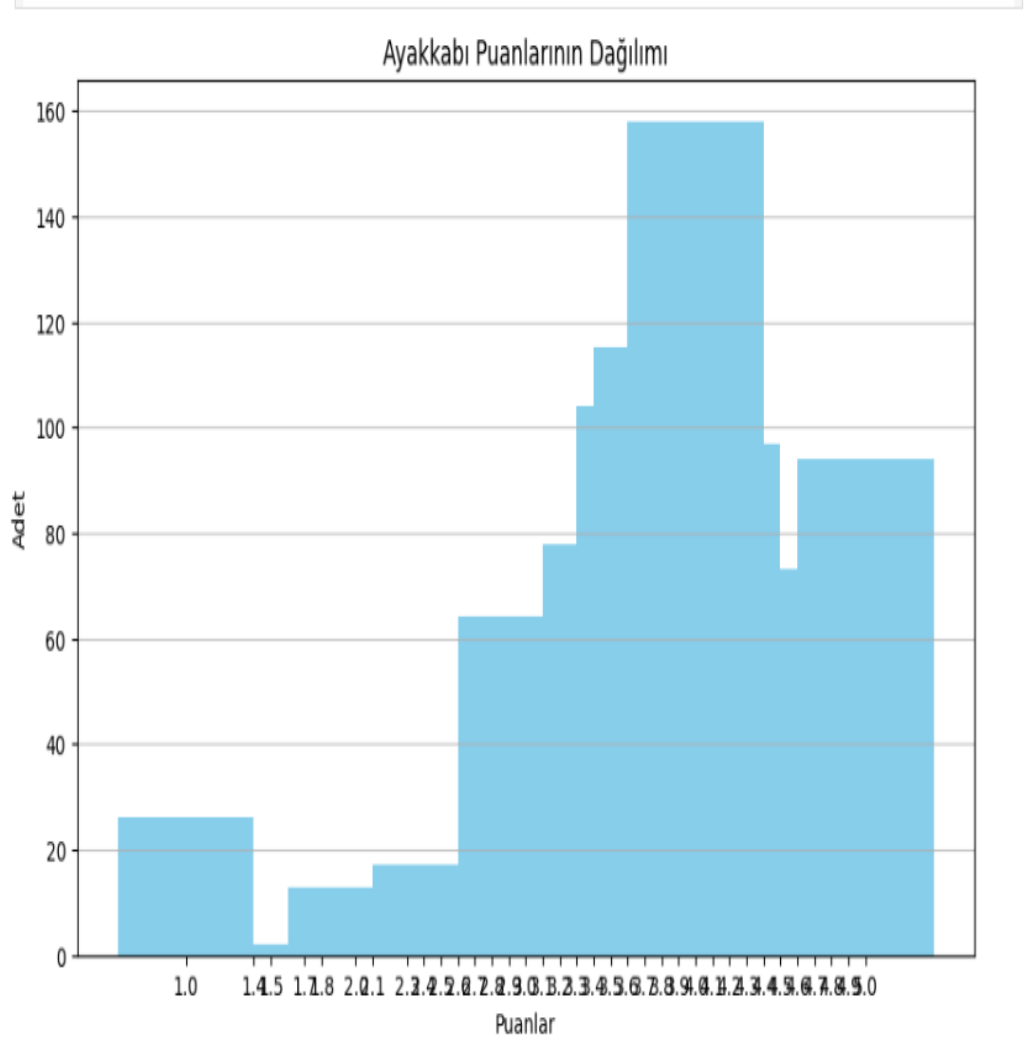
neredeyse aynı olduğunu gösteriyor. Yani bu ayakkabılar indirimsiz satılıyor.

- **İndirimli ve gerçek fiyat arasında pozitif bir korelasyon var:** Grafikteki noktalar genel olarak sol alt köşeden sağ üst köşeye doğru bir eğilim gösteriyor. Bu, gerçek fiyatı yüksek olan ayakkabıların indirimli fiyatlarının da genellikle yüksek olduğunu gösteriyor.
- **Farklı fiyat aralıklarında farklı yoğunluklar var:** Grafikte bazı bölgelerde noktalar daha yoğun, bazı bölgelerde ise daha seyrek. Bu, belirli fiyat aralıklarında daha fazla ayakkabı satıldığını gösteriyor.

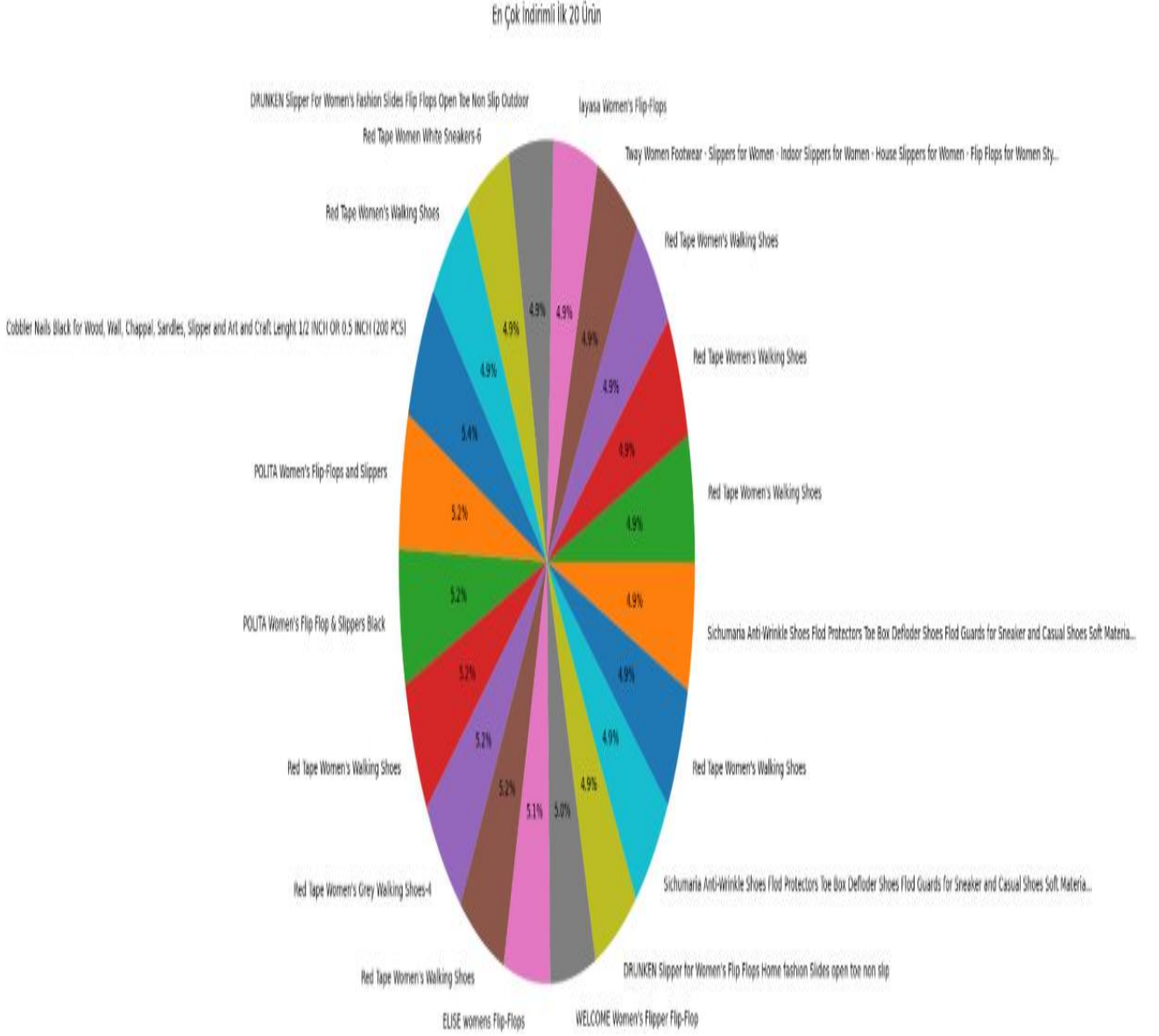
Özetle, bu grafik ayakkabı fiyatlandırması ve indirimler hakkında genel bir fikir veriyor. İndirimli fiyat ile gerçek fiyat arasındaki ilişkiyi, indirimsiz satılan ayakkabıları ve farklı fiyat aralıklarındaki yoğunlukları görselleştirerek, ayakkabı satış stratejileri hakkında bilgi edinmemizi sağlıyor.



Bu kelime bulutu, ayakkabı isimlerinde en çok kullanılan kelimeleri gösteriyor. Büyük kelimeler daha sık kullanılıyor. "Sandal", "Women", "Slipper" gibi kelimeler öne çıkıyor, yani bu kelimeleri içeren ayakkabı isimleri daha fazla. Kelime bulutu sayesinde ayakkabı isimlendirme trendlerini anlayabiliyoruz.



Bu grafik, ayakkabıların aldığı puanların dağılımını gösteriyor. Görüldüğü gibi en çok ayakkabı 3 ile 4 puan arasında almış. 1 puan alan ayakkabı sayısı oldukça az. Yani genel olarak ayakkabılar yüksek puanlar almış.



Bu pasta grafiği, **en çok indirimi olan 21 ürünü** ve bu ürünlerin indirim oranlarını gösteriyor. Her dilim bir ürünü temsil ediyor ve dilimin boyutu, o ürünün toplam indirimler içindeki payını gösteriyor.

Grafiğe bakarak şunları söyleyebiliriz:

- **İndirimler birçok ürüne dağılmış:** Hiçbir ürünün ezici bir üstünlüğü yok. İndirimler, farklı ürünler arasında nispeten dengeli bir şekilde dağılmış.
- **Bazı ürünler biraz daha öne çıkıyor:** "Drunkn Moon Slippers for Women Winter Soft Warm.." ve "Truffle Collection Women's Slippers" gibi birkaç ürün, diğerlerinden biraz daha büyük dilimlere sahip. Bu, bu ürünlerin indirim oranlarının biraz daha yüksek olduğunu gösteriyor.
- **İndirim oranları genel olarak düşük:** Çoğu dilimin üzerindeki yüzdelik değerler düşük, genellikle %5'in altında. Bu, indirimlerin genellikle küçük miktarlarda olduğunu gösteriyor.

Genel olarak, bu grafik bize şunları anlatıyor:

- İncelenen veri setinde, indirimler çok sayıda ürüne yayılmış durumda.
- Çok büyük indirimler yok, indirimler genellikle küçük miktarlarda yapılmış.
- Birkaç ürün, indirim oranları açısından biraz daha öne çıkıyor.

Bu bilgiler, ürünlerin fiyatlandırma ve indirim stratejileri hakkında fikir verebilir. Örneğin, hangi ürünlerin daha fazla indirimle satıldığı, indirimlerin genel olarak ne kadar büyük olduğu gibi sorulara cevap bulunabilir.