Python ile Birliktelik Analizi:

İş Anlama (Business Understanding):

Amaç:

Hindistan'daki 1901–2015 yıllarına ait aylık yağış verilerini kullanarak yağış miktarlarının birliktelik ilişkilerini (association rules) analiz etmek. Projenin amacı, uzun dönemli iklim verilerinde yağış miktarlarının ortak desenlerini belirlemek ve "düşük-orta-yüksek" yağış kategorileri arasında ilişki kurallarını keşfetmektir.

CRISP-DM 2: Veri Anlama (Data Understanding)

Bu aşamanın amacı, "Rainfall in India (1901–2015)" veri setinin yapısını, kapsamını ve kalitesini analiz etmektir.Verinin genel özellikleri incelenmiş, eksik değerler ve sütun türleri belirlenmiştir.Bu sayede modelleme için verinin uygunluğu değerlendirilmiştir.

Analiz Bulguları:

- Veri setinde 1901–2015 yılları arasında 36 farklı bölgeye ait yağış ölçümleri bulunmaktadır.
- `SUBDIVISION` sütunu kategorik, `YEAR` sütunu sayısal, diğer sütunlar ise sürekli (float) türündedir.
- Bazı sütunlarda eksik veriler ("NA") tespit edilmiştir. Bu değerlerin sayısı sütundan sütuna değişmektedir.
- `describe()` fonksiyonu ile yapılan özet analiz, yıllık yağış değerlerinin ortalama 100–300 mm aralığında olduğunu göstermiştir.
- Yağış miktarlarında mevsimsel farklılık gözlemlenmiştir: özellikle Haziran–Eylül aylarında yüksek değerler (muson dönemi) öne çıkmaktadır.

CRISP-DM 3: Veri Hazırlama (Data Preparation):

Bu aşamanın amacı, "Rainfall in India (1901–2015)" veri setini birliktelik analizi (Apriori algoritması) için temiz, düzenli ve kategorik hale getirmektir.Eksik değerlerin giderilmesi, veri tiplerinin dönüştürülmesi ve sayısal değişkenlerin sınıflara ayrılması işlemleri gerçekleştirilmiştir.

```
# 'NA' ifadelerini NaN'a çevir
df = df.replace("NA", pd.NA)

# Sayısal sütunları belirleme
numeric_cols = df.columns.drop(["SUBDIVISION", "YEAR"])

# Sayısal değerlere dönüştürme
df[numeric_cols] = df[numeric_cols].apply(pd.to_numeric, errors='coerce')

# Eksik değerleri medyan ile doldurma
df[numeric_cols] = df[numeric_cols].fillna(df[numeric_cols].median())
```

- Eksik değerlerin büyük kısmı medyan ile doldurulmuştur. Veri setinde artık eksik değer bulunmamaktadır.Bu adım, Apriori algoritmasının gerektirdiği eksiksiz veri yapısını sağlamıştır.

```
def categorize(value):
    if value < 100:
        return "Low"
    elif value < 300:
        return "Medium"
    else:
        return "High"

df[numeric_cols] = df[numeric_cols].applymap(categorize)
```

- Her sütundaki yağış değerleri “Low–Medium–High” kategorilerine ayrılmıştır.Bu sayede veri sürekli formdan ayrık (discrete) hale getirilmiş, Apriori algoritmasıyla çalışabilir duruma getirilmiştir.

Genel Değerlendirme:

- “NA” ifadeleri temizlenmiş, eksik veriler medyanla doldurulmuştur.
- Sayısal sütunlar, “Low”, “Medium” ve “High” kategorilerine dönüştürülmüştür.
- Gereksiz sütunlar kaldırılmış, veri analize hazır hale getirilmiştir.
- Veri artık **Apriori (Association Rule Mining)** algoritmasının giriş formatına uygundur.

CRISP-DM 4: Modelleme (Modeling)

Bu aşamada amaç, diğer aylardaki yağış miktarlarını kullanarak **Haziran ayı yağışını (JUN)** tahmin etmek ve farklı regresyon modellerinin performanslarını karşılaştırmaktır.

Python ortamında aşağıdaki adımlar uygulanmıştır:

- Veri seti **pandas** ile okunmuş, “NA” değerleri eksik olarak tanımlanmış ve tüm yağış sütunları sayısal türe dönüştürülmüştür.
- SUBDIVISION ve YEAR dışındaki sütunlar **bağımsız değişken (X)**, JUN sütunu **bağımlı değişken (y)** olarak seçilmiştir.
- Eksik değerler ilgili sütunun **medyanı** ile doldurulmuştur.
- Veri, `train_test_split` fonksiyonu ile %80 eğitim – %20 test olacak şekilde ikiye ayrılmıştır.

AutoML yaklaşımı için aşağıdaki regresyon modelleri otomatik olarak eğitilmiş ve test seti üzerinde değerlendirilmiştir:

- **LinearRegression**
- **DecisionTreeRegressor**
- **RandomForestRegressor**
- **GradientBoostingRegressor**
- **SVR (RBF kernel)**

Her model için **R² (belirleme katsayısı)** ve **RMSE (Root Mean Squared Error)** metrikleri hesaplanmış ve sonuçlar aşağıdaki gibi elde edilmiştir:

Model	R²	RMSE
LinearRegression	0.9993	76.62
RandomForest	0.8925	76.62
GradientBoosting	0.8884	76.62
DecisionTree	0.7777	76.62
SVR	0.7285	76.62

Bu tabloya göre **en yüksek R² değeri** LinearRegression modelinde elde edilmiştir.
Dolayısıyla, AutoML süreci sonunda **en iyi model olarak Linear Regression seçilmiştir.**

CRISP-DM 5: Değerlendirme (Evaluation):

- **Linear Regression modeli**, $R^2 \approx 0.9993$ ile Haziran yağışlarının neredeyse tamamını açıklamaktadır.
  Bu, diğer aylardaki yağış değerleri ile Haziran ayı yağışı arasında **çok güçlü doğrusal bir ilişki** olduğunu göstermektedir.

- RandomForest ve GradientBoosting modelleri de $R^2 \approx 0.89$ civarı ile kabul edilebilir performans vermiş, ancak Linear Regression kadar başarılı olamamıştır.
- DecisionTree ve SVR modellerinin R^2 değerleri daha düşüktür; bu modeller veri setindeki doğrusal ilişkiyi Linear Regression kadar iyi yakalayamamıştır.

Genel değerlendirme:

AutoML sürecinde denenen beş farklı regresyon modeli arasında **Linear Regression**, en yüksek R^2 değeri ile en başarılı model olarak belirlenmiştir. Bu sonuç, Haziran ayı yağış miktarının diğer ayların yağışlarına göre oldukça iyi tahmin edilebildiğini ve veri seti içerisinde güçlü bir doğrusal yapı bulunduğunu göstermektedir. Model, birliktelik analizinde gözlenen yağış desenlerini sayısal olarak da doğrulamakta ve iklim verilerinin öngörülmesinde kullanılabilir bir tahminleyici sunmaktadır.

-