

## **CRISP-DM 1: İş Anlama (Business Understanding):**

Bu aşama, veri madenciliği sürecinin temelidir. Henüz KNIME'a geçmeden önce problemi tanımlar, hedefi belirler ve veriyle ne öğrenmek istediğimizi açıklarız.

### **Amaç (Business Objective):**

Hedefimiz:

Hindistan'da 1901–2015 yılları arasında farklı bölgelerdeki yağış miktarlarının tarımsal veya mevsimsel desenlerle birlikte nasıl gerçekleştiğini anlamak.

Başka bir ifadeyle:

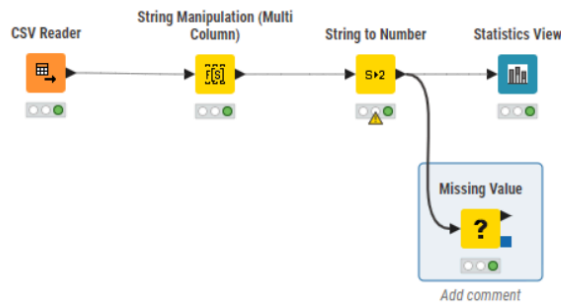
- “Yağış miktarları arasında hangi bölgeler veya aylar birlikte yüksek/düşük yağış eğilimi gösteriyor?”
- “Bazı bölgelerde yüksek yağış varsa, başka hangi bölgelerde de yüksek olma olasılığı artıyor?”  
gibi birliktelik kuralları (association rules) elde edeceğiz.

## **CRISP-DM 2: Veri Anlama (Data Understanding):**

Bu aşamanın amacı, “Rainfall in India (1901–2015)” veri setinin yapısını, içeriğini ve kalitesini inceleyerek veri hazırlama sürecine temel oluşturmaktır. Verinin genel özellikleri, eksik değerler, değişken türleri ve istatistiksel özetleri analiz edilmiştir.

### **Veri Seti Bilgisi**

- Kaynak: Kaggle – *Rainfall in India (1901–2015)*
- Kapsam: 1901–2015 yılları arasında Hindistan'daki 36 farklı bölgeye ait aylık yağış miktarları.
- Değişkenler:
  - SUBDIVISION: Bölge adı (kategorik)
  - YEAR: Yıl bilgisi (sayısal)
  - JAN – DEC: Her aya ait ortalama yağış miktarı (mm cinsinden)
  - ANNUAL, Jan-Feb, Mar-May, Jun-Sep, Oct-Dec: Mevsimsel ve yıllık toplam yağış sütunları (mm)



Node	İşlev	CRISP-DM 2'deki Görevi
CSV Reader	Veri setini (Rainfall in India 1901–2015.csv) KNIME'a aktardın.	Veri toplama ve ilk gözlem (Data Collection).
String Manipulation (Multi Column)	"NA" ifadelerini boş (?) değere dönüştürdün.	Eksik/hatalı veriyi belirleme ve temizleme.
String to Number	Sayısal sütunları Double tipe çevirdin.	Veri tiplerini düzenleme (numeric–categorical ayrımı).
Statistics View	Min, max, ortalama, eksik değer sayısı gibi istatistikleri görüntüledin.	Verinin yapısını ve dağılımını anlama.
Missing Value	Eksik değerlerin oranını ve dağılımını gözlemledin (Do nothing modunda).	Veri kalitesini ölçme, eksik veri analizi.
Column Filter	Gereksiz sütunları (ör. YEAR) çıkararak sadeleştirmeye başladın.	Veri sadeleştirmeye hazırlık (bir sonraki aşamaya geçiş).

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

Bu aşamanın amacı, "Rainfall in India (1901–2015)" veri setini birliktelik analizi (Apriori algoritması) için temiz, kategorik ve analize hazır hale getirmektir.

Veri setinde eksik değerler düzeltilmiş, gereksiz sütunlar çıkarılmış ve sayısal sütunlar kategorik değerlere dönüştürülmüştür.

#### **Sayısal Verilerin Kategorik Hale Getirilmesi:**

**Node:** Numeric Binner

Yağış değerleri 3 aralıkta sınıflandırılmıştır:

Kategori	Aralık (mm)	Açıklama
Low	0 – 100	Düşük yağış
Medium	100 – 300	Orta yağış
High	300+	Yüksek yağış

## CRISP-DM 4 – MODELLEME (Modeling):

Bu aşamanın amacı, “Rainfall in India (1901–2015)” veri setinden çıkarılan kategorik değerler (Low–Medium–High) üzerinden Apriori algoritması ile yağış düzeyleri arasındaki genel birliktelikleri belirlemektir.



### Veri Girişi

Bir önceki aşamada hazırlanan kategorik veri (Low, Medium, High) sütunları,

**Create Collection Column** node'u kullanılarak “Transactions” adıyla tek bir koleksiyon sütununa dönüştürülmüştür.

Her satır, bir bölgenin aylık ve mevsimsel yağış kategorilerini temsil eden bir liste olarak kaydedilmiştir.

### Model Kurulumu

- **Algoritma:** Association Rule Learner (Borgelt)
- **Transaction Column:** Transactions
- **Minimum Support:** 0.9
- **Minimum Confidence:** 0.8
- **Itemset Type:** CLOSED
- **Maximum Itemset Length:** 10

Bu parametrelerle Apriori algoritması uygulanarak en sık görülen birliktelik kuralları çıkarılmıştır.

Kural (Rule)	Support	Confidence	Lift
[Medium] → [High]	0.96	0.978	1.001
[High] → [Medium]	0.96	0.983	1.001

- Bu durum, genel olarak Hindistan'ın yağışlı bölgelerinde “yüksek” ve “orta” seviyedeki yağışların benzer dönemlerde gerçekleştiğini, yani **muson mevsimlerinde yüksek korelasyon** bulunduğunu ortaya koymaktadır. Lift değeri 1'e çok yakın olduğu için kurallar güçlü korelasyon göstermemektedir; ancak verinin genel yapısında **yüksek yağışların baskın** olduğu görülmektedir.

## **CRISP-DM 5: Değerlendirme (Evaluation):**

### **Modelin Değerlendirilmesi:**

<b>Değerlendirme Kriteri</b>	<b>Sonuç</b>	<b>Açıklama</b>
Doğruluk (Confidence)	Yüksek (>%95)	Kuralların doğruluk oranı güçlüdür.
Destek (Support)	Çok yüksek (%96)	Kurallar veri setinin büyük bölümünde geçerlidir.
Lift	1.0 civarında	İlişkiler güçlü korelasyon göstermemektedir.
Modelin genellenebilirliği	Orta	Yüksek support nedeniyle kurallar geneldir.

## **CRISP-DM 6 – Dağıtım (Deployment):**

### **Amaç**

Bu aşamanın amacı, oluşturulan birliktelik analizi modelinden elde edilen sonuçları karar vericilere sunmak, modelin çıktılarının uygulanabilirliğini değerlendirmek ve gelecekteki benzer projeler için bir yol haritası oluşturmaktır.

### **Gelecek Çalışmalar için Öneriler**

1. **Support değeri düşürülerek** (ör. 0.2) daha fazla kural çıkarılabilir.
2. **Sütun adları eklenerek** ("JAN=High", "FEB=Medium" gibi) ay bazlı kurallar tespit edilebilir.
3. Aynı metodoloji farklı bölgelerin (örneğin Kuzey Hindistan, Güney Hindistan) alt kümelerine uygulanarak karşılaştırmalı analiz yapılabilir.
4. Model sonuçları, zaman serisi analizleriyle birleştirilerek gelecekteki yağış tahminlerine destek sağlayabilir.