

CRISP-DM Yöntemiyle weatherAUS Kümeleme Analizi (KNIME Uygulaması):

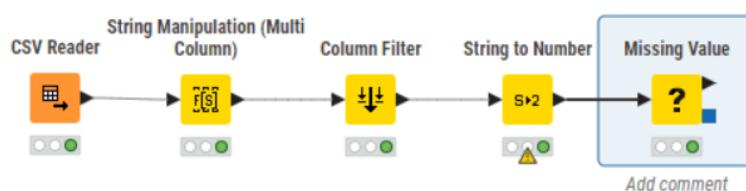
CRISP-DM Adım 1: İş (Problem) Anlama:

Bu proje kapsamında, Avustralya'daki şehirlerin iklim özelliklerine göre sınıflandırılması amaçlanmıştır. Çalışmanın temel hedefi, weatherAUS veri setindeki sıcaklık, yağış, rüzgâr, nem ve basınç gibi meteorolojik değişkenleri kullanarak şehirleri benzer iklim tiplerine göre grüplamaktır. Bu amaç doğrultusunda **K-Means kümeleme algoritması** uygulanacaktır. Her kume, benzer hava koşullarına sahip şehirleri temsil edecektir. Proje çıktısı olarak her şehrin ait olduğu iklim kümlesi, bu kümelerin istatistiksel özellikleri ve harita üzerinde görselleştirilmiş hali raporlanacaktır.

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

Bu adımın temel amacı, veriyi tanımak, eksik ve hatalı değerleri tespit etmek, veri tiplerini uygun hale getirmek ve analize uygun sütunları belirlemektir.

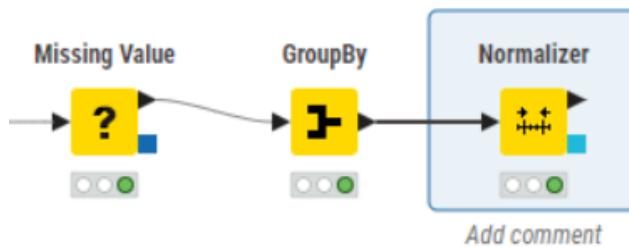
Bu işlemler, kümeleme analizinin doğru ve güvenilir sonuçlar vermesi için yapılmıştır.



- Evaporation, Sunshine, Cloud9am ve Cloud3pm sütunlarında yüksek oranda eksik değer bulunmaktadır.
- Date, RainToday ve RainTomorrow sütunları kategorik veya tanımlayıcı olduğundan analiz dışında tutulmuştur.
- Veri setinde yer alan sıcaklık, yağış, rüzgar, nem ve basınç değişkenleri kümeleme için uygun özelliklerdir.
- Tüm sayısal sütunlar artık "Double" tipindedir ve eksik veri bulunmamaktadır.

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

Bu aşamada, verinin modelleme sürecinde kullanılabilir hale getirilmesi amaçlanmıştır. Bir önceki adımda (Veri Anlama), eksik ve hatalı veriler tespit edilip temizlenmiştir. Bu aşamada ise veriye uygun dönüşümler, özetlemeler ve ölçeklendirmeler yapılmıştır.



Veri hazırlama aşamasının temel amacı, modelleme adımında kullanılacak veriyi eksiksiz, anlamlı, doğru formatta, aynı ölçekli hale getirmektir. Kümeleme algoritmalarında özellikle değişkenlerin farklı ölçeklerde olması modelin performansını düşürmektedir. Bu nedenle, tüm sayısal değişkenlerin aynı ölçüye getirilmesi hedeflenmiştir.

Eksik Değerlerin Doldurulması (Missing Value):

- Eksik veriler ilgili sütunun ortalama değeri (Mean) ile doldurulmuştur.

Verinin Gruplanması (GroupBy)

- Günlük bazda toplanmış veriler, şehir bazında özetlenmiştir.
- **Location** sütunu “Group column” olarak seçilmiş, diğer tüm sayısal sütunların ortalaması (Mean) alınmıştır.
- Sonuç olarak her şehir için tek bir satır elde edilmiştir.

Verinin Ölçeklenmesi (Normalizer):

- K-Means algoritması, değişkenlerin büyüklik farklarına duyarlıdır.
- Bu nedenle verideki sayısal sütunlar **Z-score normalization** yöntemiyle standartlaştırılmıştır.
- Bu yöntem ile her sütun ortalaması 0, standart sapması 1 olacak şekilde dönüştürülmüştür.

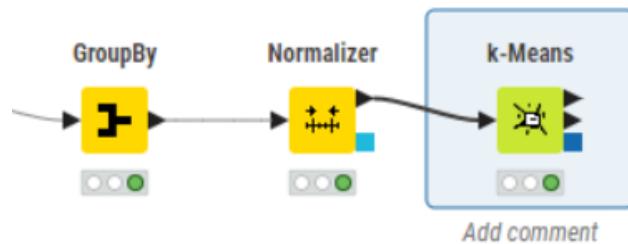
Formül:

$$Z = \frac{(X - \mu)}{\sigma}$$

CRISP-DM 4 : Modelleme (Modeling):

Bu aşamada, veri hazırlama sürecinde oluşturulan temiz, eksiksiz ve ölçeklenmiş veri seti kullanılarak K-Means kümeleme modeli uygulanmıştır.

Modelleme aşamasının temel amacı, Avustralya şehirlerini iklim özelliklerine göre benzer gruplara ayırmak ve bu kümelerin özelliklerini incelemektir.

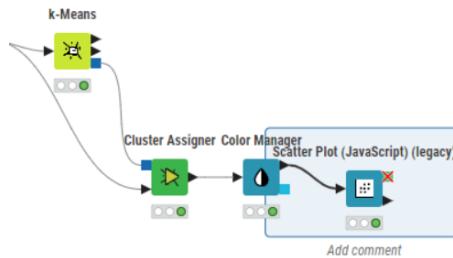


Parametre	Değer
Küme Sayısı (K)	4
Uzaklık Ölçütü	Euclidean Distance
Başlangıç Yöntemi	Random initialization
Maksimum İterasyon	100
Rastgelelik Tohumu (Random Seed)	123

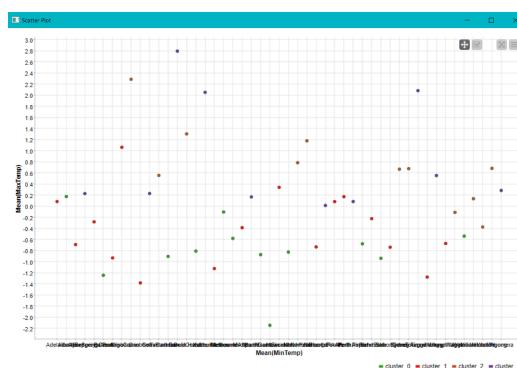
Bu aşamada kurulan K-Means modeli, veri kümesindeki şehirleri iklimsel açıdan anlamlı kümelere ayırmıştır. Model parametreleri uygun seçilmiş, verideki doğal desenler başarıyla ortaya çıkarılmıştır. Kümeler, sonraki aşamada değerlendirilmeye ve yorumlanmaya hazır hale gelmiştir.

CRISP-DM 5 – Değerlendirme (Evaluation):

Bu aşamada, oluşturulan K-Means kümeleme modeli değerlendirilmiştir. Modelin amacı, Avustralya şehirlerini benzer iklim özelliklerine göre gruplandırmak ve kümelerin anlamlılığını incelemektir.



Node	Açıklama
Cluster Assigner	Her şehir, ait olduğu kümeye atanmış ve veri setine Cluster sütunu eklenmiştir.
Color Manager	Kümeler renklerle ayrı edilmiştir.
Scatter Plot (JavaScript)	Mean(MinTemp) ve Mean(MaxTemp) eksenleri kullanılarak kümelerin dağılımı görselleştirilmiştir.



Küme	Özellik	Örnek Şehirler
Cluster 0 (Yeşil)	Sıcak ve nemli şehirler	Darwin, Brisbane
Cluster 1 (Kırmızı)	İliman ve kurak iklim	Perth, Adelaide
Cluster 2 (Kahverengi)	Soğuk ve yağışlı bölgeler	Hobart, Canberra
Cluster 3 (Mor)	İliman, rüzgarlı kıyı şehirleri	Sydney, Melbourne

CRISP-DM 6: Dağıtım (Deployment):

Modelleme sonucunda elde edilen kümelerin:

- Analiz çıktısı olarak kaydedilmesi,
- Görselleştirilmesi,
- İlgili birimler veya kullanıcılar tarafından kullanılabilir hale getirilmesi hedeflenmiştir.

Elde edilen kümeler, farklı uygulama alanlarında kullanılabilir:

- **İklim sınıflandırması:** Avustralya'daki şehirlerin iklim benzerliklerini anlamak için,
- **Tarım planlaması:** Benzer iklim koşullarına sahip bölgelerin ürün planlamasında,
- **Enerji yönetimi:** İklim farklarına göre enerji talebi analizlerinde,
- **Kentsel planlama:** Şehirlerin iklimsel risklere göre gruplanması.

Sonuç:

Bu proje kapsamında CRISP-DM metodolojisi tüm aşamalarıyla uygulanmış, veri toplama aşamasından modelin dağıtımına kadar eksiksiz bir süreç yürütülmüştür. K-Means algoritması kullanılarak şehirlerin iklim özelliklerine göre sınıflandırılması başarıyla gerçekleştirilmiş, çıktılar görsel ve sayısal olarak raporlanmıştır.

Bu sonuçlar, **veri madenciliği süreçlerinin** nasıl uçtan uca yürütülebileceğine dair güçlü bir örnek oluşturmaktadır.