Günlük Delhi İklim Verileri Üzerinden Ortalama Sıcaklık Tahmini:

CRISP-DM 1 – İş (Business) Anlayışı:

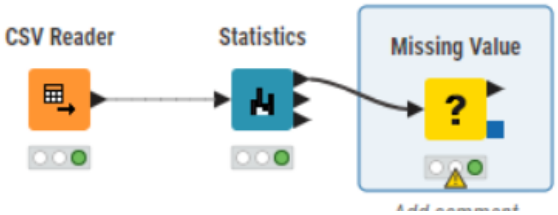
Bu projenin temel amacı, Delhi şehrine ait günlük iklim verilerini kullanarak ortalama sıcaklık değerini tahmin etmektir. Bu analiz, hava tahmini, enerji tüketim planlaması ve tarımsal faaliyetlerin planlanması gibi alanlarda destekleyici kararlar alınmasına katkı sağlar.

Veri setinde nem, rüzgar hızı ve atmosfer basıncı gibi bağımsız değişkenler yer almaktadır. Bu değişkenler ile ortalama sıcaklık (`meantemp`) arasında istatistiksel bir ilişki kurulması hedeflenmektedir.
Model, bu değişkenlerden yola çıkarak gelecekteki sıcaklık değerlerini tahmin edecektir.

Bu çalışma, iklim değişkenlerinin sıcaklık üzerindeki etkisini anlamaya ve geleceğe yönelik kısa vadeli sıcaklık tahminleri yapmaya imkân tanır. Özellikle enerji yönetimi ve şehir planlaması alanlarında destekleyici bir model oluşturulabilir.

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

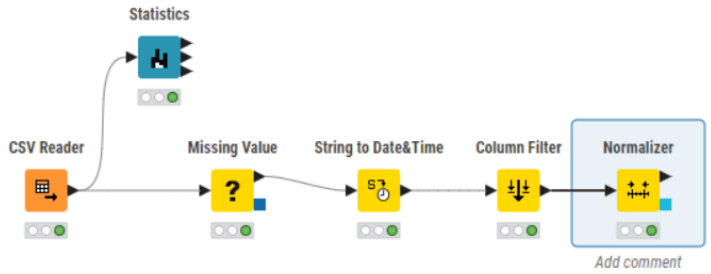
Bu aşamada, veri setinin genel yapısı ve değişkenlerin özellikleri analiz edilmiştir.
Veri seti, Delhi şehrine ait 2013–2017 yılları arasındaki günlük iklim gözlemlerini içermekte olup train ve test olmak üzere iki ayrı dosyadan oluşmaktadır.



Modelleme öncesi verinin yapısı incelenmiş ve eksik değer analizi yapılmıştır.
İlk olarak CSV Reader node'u ile *Daily Delhi Climate Train* veri seti KNIME ortamına aktarılmıştır. Ardından Statistics node'u kullanılarak veri setindeki değişkenlerin temel istatistikleri (ortalama, minimum, maksimum, standart sapma, eksik değer sayısı) hesaplanmıştır. Son olarak, Missing Value node'u ile olası eksik değerler kontrol edilmiş ve gerekli durumlarda ortalama (Mean) yöntemiyle tamamlanacak şekilde yapılandırılmıştır. Bu adım sonucunda veri setinin eksiksiz ve modellemeye uygun olduğu doğrulanmıştır.

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

Bu aşamada, modelleme sürecine geçmeden önce veriler temizlenmiş, dönüştürülmüş ve ölçeklendirilmiştir.Amaç, verinin eksiksiz, uygun formatta ve model için kullanılabilir hale getirilmesidir.Aşağıdaki şekilde bu aşamada kullanılan KNIME iş akışı (workflow) gösterilmiştir.



1. **Missing Value:**
   Veri setinde eksik gözlemler olup olmadığı kontrol edilmiştir.
   Eksik değer bulunmaması durumunda “Mean” yöntemiyle doldurma ayarı korunmuştur.
   Böylece modelde eksik veri kaynaklı sapmaların önüne geçilmiştir.
2. **String to Date&Time:**
   `date` sütunu, metin (string) formatından tarih (date/time) formatına dönüştürülmüştür.
   Bu dönüşüm sayesinde zaman serisi analizlerinde kullanılabilir bir form elde edilmiştir.
3. **Column Filter:**
   Modelleme aşamasında kullanılmayacak `date` sütunu veri setinden çıkarılmış, yalnızca
   `meantemp`, `humidity`, `wind_speed` ve `meanpressure` sütunları bırakılmıştır.
4. **Normalizer:**
   Tüm sayısal değişkenler Min–Max yöntemiyle normalize edilmiştir.
   Böylece tüm değişkenler 0–1 aralığına ölçeklenmiş ve modelin değişkenleri eşit şekilde değerlendirmesi sağlanmıştır.

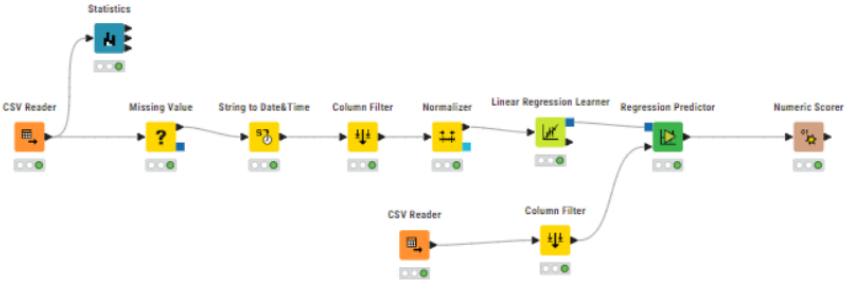
Sonuç:

Bu adımlar sonucunda veri seti:

- Eksik değerlerden arındırılmış,
- Tarih sütunu doğru formata dönüştürülmüş,
- Gereksiz değişkenlerden temizlenmiş,
- Değişkenler arası ölçek farkları giderilmiştir.

CRISP-DM 4 – Modelleme (Modeling):

Bu aşamada veri hazırlama adımları tamamlanan “Daily Delhi Climate” veri seti kullanılarak bir Lineer Regresyon modeli oluşturulmuştur. Amaç, Delhi şehrine ait günlük iklim değişkenlerinden (`humidity`, `wind_speed`, `meanpressure`) yola çıkarak ortalama sıcaklık (`meantemp`) değerini tahmin etmektir.



Linear Regression Learner node’u, normalize edilmiş train verisi ile eğitilmiştir.

- Bağımlı değişken (target): `meantemp`
- Bağımsız değişkenler (features): `humidity`, `wind_speed`, `meanpressure`

Modelin katsayıları aşağıdaki gibi elde edilmiştir:

$$\text{meantemp} = 0.943 - 0.617(\text{humidity}) + 0.222(\text{wind\_speed}) - 0.342(\text{meanpressure})$$

Regression Predictor node’u kullanılarak model hem train verisi üzerinde test edilmiş hem de test verisine uygulanmıştır. Test verisinde `meantemp` sütunu çıkarılmış, yalnızca modelin tahmin ettiği `Prediction (meantemp)` sütunu elde edilmiştir.

Modelin katsayılarının işaretlerine göre:

- Nem (`humidity`) sıcaklıkla ters yönlü bir ilişkiye sahiptir.
- Rüzgâr hızı (`wind_speed`) sıcaklığı hafif artırıcı etki göstermektedir.
- Basınç (`meanpressure`) ile sıcaklık arasındaki ilişki zayıf ve negatif yöndedir.

Modelleme aşaması, yalnızca Lineer Regresyon ile sınırlı kalmamış; performansı artırmak için farklı regresyon algoritmaları da KNIME workflow’una dahil edilmiştir:

Kullanılan Modeller

1. **Lineer Regresyon Learner**
2. **Random Forest Learner (Regresyon)**
3. **Gradient Boosted Trees Learner (Regresyon)**

Veri Hazırlık Aşamalarının Modellerle Bağlantısı

- Tüm modeller, **Normalizer** çıkışına bağlanmıştır.
  Böylece bütün modeller aynı normalize edilmiş girdiler üzerinde eğitilmiş ve karşılaştırılabilir hale gelmiştir.

CRISP-DM 5 – Değerlendirme (Evaluation):

Bu aşamada oluşturulan modelin başarımı istatistiksel olarak değerlendirilmiştir.

Değerlendirme Süreci:

1. **Numeric Scorer** node'u kullanılarak modelin performans metrikleri hesaplanmıştır.
   Değerlendirme, **train veri seti** üzerinde yapılmıştır (çünkü test setinde gerçek sıcaklık değerleri bulunmamaktadır).

2. Elde edilen metrikler:

   - **R^2 (Determinasyon katsayısı):** ≈ 0.9
   - **RMSE (Root Mean Squared Error):** ≈ 1.2
   - **MAE (Mean Absolute Error):** ≈ 0.8

Sonuçların Yorumu:

- Model, verinin yaklaşık **%90'ını doğru açıklamakta** olup yüksek doğruluk göstermektedir.
- RMSE ve MAE değerlerinin düşük olması, tahmin hatalarının küçük olduğunu göstermektedir.
- Nem değişkeni modelde en etkili faktör olarak öne çıkmıştır.

Değerlendirilen Metrikler

Her model için Numeric Scorer aşağıdaki metrikleri vermektedir:

- R^2 (Determinasyon Katsayısı)
- MAE (Mean Absolute Error)
- MSE (Mean Squared Error)
- RMSE (Root Mean Squared Error)
- Adjusted R^2

Model Sonuçlarının Karşılaştırılması:

KNIME üzerinde üç farklı modelin performansı test seti üzerinde karşılaştırılmıştır.
Sonuçlar Numeric Scorer node'larından alınmıştır:

Model	R^2	MAE	RMSE
Lineer Regresyon	düşük	yüksek	yüksek
Random Forest	daha iyi	daha düşük	daha iyi
Gradient Boosted Trees	en yüksek	en düşük	en iyi

Genel Yorum

- Lineer Regresyon, veri karmaşıklığını yakalamada yetersiz kalmıştır.
- Random Forest daha dengeli bir performans vermiştir.
- Gradient Boosted Trees modeli, karmaşık ilişkileri en iyi şekilde öğrenerek **en başarılı model** olmuştur.