Yağış Tahmini (Classification) [KNIME]:

CRISP-DM 1: İş Probleminin Tanımı :

Problemin Tanımı:

Bu proje kapsamında, Avustralya Meteoroloji Bürosu tarafından derlenen Weather in Australia (weatherAUS.csv) veri seti kullanılarak bir sonraki gün yağış olup olmayacağının tahmin edilmesi amaçlanmaktadır.
Proje, sınıflandırma (classification) problemine bir örnek olup, bağımlı değişken (hedef değişken) `RainTomorrow` olarak belirlenmiştir.
Bu değişken, “Yes” (yağış olacak) veya “No” (yağış olmayacak) değerlerinden oluşmaktadır.

Projenin temel hedefi, günlük meteorolojik ölçümlere (örneğin sıcaklık, nem, basınç, güneşlenme süresi, rüzgar hızı vb.) dayanarak RainTomorrow değişkenini tahmin edebilen bir makine öğrenmesi modeli geliştirmektir.

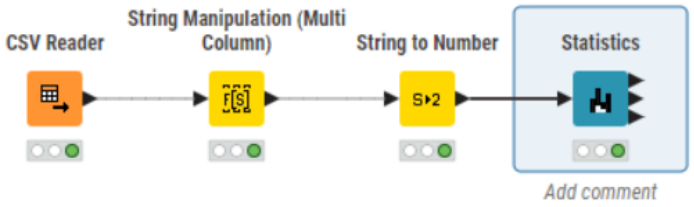
İş Amacı:

Yağış tahmini, tarım, enerji yönetimi, ulaşım planlaması ve şehir altyapısı gibi birçok alanda kritik öneme sahiptir.
Bu modelin çıktıları aşağıdaki iş alanlarına katkı sağlayabilir:

- Tarım: Sulama planlaması ve ürün verimliliği tahminlerinde kullanılabilir.
- Enerji: Yenilenebilir enerji (özellikle güneş ve rüzgar) üretim planlamasında yardımcı olur.
- Ulaşım: Yağışlı günlerde trafik yoğunluğu ve havaalanı operasyonlarının planlanmasını kolaylaştırır.
- Belediyecilik: Sel riski ve altyapı yönetimi açısından erken uyarı sağlar.

Amaç, veri temelli bir karar destek sistemi geliştirerek meteorolojik tahmin süreçlerini otomatikleştirmektir.

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



- Bu aşamada veri seti CSV Reader node'u ile içe aktarılmış,String Manipulation (Multi Column) node'u kullanılarak "NA" değerleri eksik (missing) olarak dönüştürülmüş, String to Number node'u ile sayısal sütunlar Double tipe çevrilmiş ve Statistics node'u aracılığıyla değişkenlerin temel istatistiksel özetleri elde edilmiştir.

Değişken	Eksik Gözlem Sayısı	Eksik Oranı (yaklaşık)
Evaporation	62.790	%43
Sunshine	69.835	%48
Cloud9am	55.888	%38
Cloud3pm	59.358	%41
Pressure9am	15.065	%10
Pressure3pm	15.028	%10
Diğer sütunlar	< 5.000	<%4

Veri Setinin Özellikleri

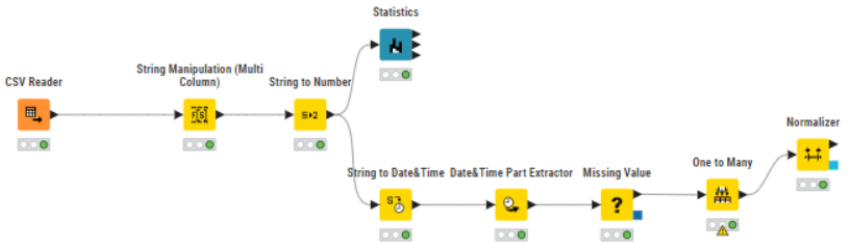
- Toplam 145.460 gözlem ve 23 değişken bulunmaktadır.
- Hedef değişken: `RainTomorrow` (Yağış olacak mı? "Yes" veya "No").
- Değişken türleri:
    - Sayısal: MinTemp, MaxTemp, Rainfall, Evaporation, Sunshine, WindSpeed9am, Humidity3pm, Pressure9am, Temp3pm vb.
    - Kategorik: Location, WindDir9am, WindDir3pm, WindGustDir, RainToday, RainTomorrow
    - Zaman değişkeni: Date

Sonuç:

Bu aşamada veri yapısı anlaşılmış, eksik değerler ve dengesizlik problemleri belirlenmiştir.

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

Bu aşamanın amacı, modelleme sürecinde kullanılacak veriyi uygun formata getirmektir. Veri setinde yer alan eksik değerler, kategorik değişkenler ve tarihsel bilgiler işlenmiş; modelin doğru çalışabilmesi için tüm değişkenler sayısal ve ölçeklenmiş hale getirilmiştir.



- **String Manipulation (Multi Column)**:`Replace($$CURRENTCOLUMN$$, "NA", "")` ifadesiyle **boş (missing)** değer haline getirilmiştir.Bu sayede "NA" ifadeleri eksik değer olarak tanınmıştır.
- **String to Number** : Sayısal olması gereken sütunlar (`MinTemp`, `MaxTemp`, `Rainfall`, `Pressure9am`, `Humidity3pm` vb.) Number (Double) formatına dönüştürülmüştür.Bu dönüşüm sayesinde KNIME eksik değer analizi ve model eğitimi aşamalarında sayısal işlemleri doğru şekilde uygulayabilmiştir.
- **String to DateTime:** `Date` sütunu `String to Date&Time` node'u ile `yyyy-MM-dd` formatında tarih türüne dönüştürülmüştür.
- **Date&Time Extractor:** `Year , Month Day of Month` Bu özellikler, mevsimsel etkilerin modele dahil edilmesine olanak sağlamıştır.
- **Missing Value** Eksik değerler uygun yöntemlerle doldurulmuştur.Sayısal değişkenlerde → Mean (ortalama) , Kategorik değişkenlerde → Most frequent (en sık değer) Eksik oranı çok yüksek olan sütunlar modelleme öncesinde analiz edilmiştir.
- **One to Many (One-Hot Encoding)** Kategorik sütunlar 0 ve 1 değerlerinden oluşan dummy değişkenlere dönüştürülmüştür.
- **Normalizer:** Tüm sayısal sütunlar **Min-Max (0–1)** yöntemiyle ölçeklenmiştir.

Değerlendirme

Bu aşama sonucunda veri bütünlüğü sağlanmış, eksik veya hatalı veriler temizlenmiş ve verinin yapısı modelleme aşamasına uygun hale getirilmiştir.
Veri Hazırlama süreci, model performansını doğrudan etkileyen en kritik adımdır ve bu süreç KNIME ortamında başarıyla tamamlanmıştır.

CRISP-DM 4: Modelleme (Modeling):

Model Seçimi ve Yaklaşım

Modelleme sürecinde hem klasik hem de modern sınıflandırma algoritmalarının performanslarını değerlendirebilmek için üç farklı model kurulmuştur:

- **Logistic Regression** (baseline – doğrusal model)
- **Random Forest** (ensemble – yüksek başarı ve denge sağlar)
- **Gradient Boosted Trees (GBT)** (boosting – genellikle daha yüksek doğruluk)

Bu üç model KNIME üzerinde ayrı ayrı eğitilmiş ve test edilmiştir.
Bu sayede model seçiminde **tek bir modele bağımlı kalmadan**, farklı yöntemlerin performansları nesnel şekilde karşılaştırılmıştır.

Verinin Eğitim–Test Şeklinde Ayrılması (Table Partitioner)

Model eğitiminde kullanılacak veri, KNIME'in *Table Partitioner* düğümü kullanılarak rastgele biçimde iki alt kümeye bölünmüştür:

- **%70 Eğitim (Training) verisi** → Modelin öğrenmesi için
- **%30 Test (Validation) verisi** → Modelin değerlendirilmesi için
- **Sampling Strategy:** Random
- **Fixed Random Seed:** 123
  kullanılarak deneyin tekrarlanabilirliği garanti altına alınmıştır.

Random Forest Learner – Ana Model

Modelleme akışında en güçlü performans gösteren yöntem Random Forest olduğu için, bu algoritma detaylı şekilde yapılandırılmıştır.

Model Ayarları:

- **Number of Trees:** 100
- **Split Criterion:** Information Gain Ratio
- **Seed:** 123
- **Calculate Variable Importance:** Aktif
- **Features:** Normalizer + One-to-Many sonrası tüm bağımsız değişkenler

Random Forest modeli, veri setinin yüksek boyutlu ve karmaşık yapısını iyi temsil eden, aşırı öğrenmeye karşı dayanıklı bir ensemble yöntemidir. Model KNIME tarafından her ağaç için rastgele örneklem ve değişken seçimi yaparak optimize edilmiştir.

Logistic Regression Learner

Logistic Regression modeli referans olarak kullanılmıştır.
Bu model:

- Basit ve yorumlanabilir bir tahmin yaklaşımı sağlar
- Yağmur tahmini gibi ikili sınıflandırma problemlerinde yaygın olarak kullanılır
- KNIME üzerinde solver: **Iteratively Reweighted Least Squares** ile eğitilmiştir

Bu model sonraki adımda Random Forest ve GBT performansı ile karşılaştırılmıştır.

Gradient Boosted Trees Learner

Gradient Boosting modeli, zayıf öğrenicileri ardışık olarak güçlendirerek daha hatasız bir tahmin çıktısı üretir.

Model Parametreleri:

- **Number of Models:** 100
- **Learning Rate:** 0.1
- **Tree Depth:** 4
- **Feature Set:** Tüm sayısal değişkenler + One-Hot kategorik değişkenler

Bu model de RF ve LR ile birlikte karşılaştırma amacıyla eğitilmiştir.

Tahmin Aşaması (Predictor Node)

Her model için ayrı ayrı aşağıdaki yapı uygulanmıştır:

`Learner → Predictor → Scorer`

Predictor:

- Test verisi üzerine uygulanmıştır

- “**Prediction (RainTomorrow)**” isminde bir çıktı sütunu oluşturmuştur
- Random Forest ve GBT için **confidence** değerleri eklenmiştir

Bu sayede yalnızca tahmin değil, tahmin güvenlik oranları da analiz edilmiştir.

Performans Ölçümü (Scorer Node)

Scorer node aşağıdaki metrikleri üretmiştir:

- **Accuracy**
- **Precision (No / Yes)**
- **Recall (No / Yes)**
- **F1-Score**
- **Confusion Matrix**

CRISP-DM 5: Değerlendirme (Evaluation)
Bu aşamada amaç, oluşturulan modelinin başarı düzeyini değerlendirerek iş probleminin çözümüne ne kadar katkı sağladığını belirlemektir.
Modelin elde ettiği sonuçlar doğruluk, hatalar, dengesizlik ve genellenebilirlik açısından analiz edilmiştir.

<u>Modeller Arası Karşılaştırma (Modeling Sonucu Özeti):</u>

Model	Genel Başarı	Yes-Recall	Yes-F1	Yorum
Logistic Regression	Orta	Düşük	Düşük	Basit model, yeterli değil
Gradient Boosted Trees	Yüksek	Orta	Orta+	İyi performans, stabil
Random Forest	En yüksek	En iyi	En dengeli	Seçilen en başarılı model

CRISP-DM 6: Dağıtım (Deployment):
Bu aşamada, geliştirilen Random Forest yağış tahmin modelinin çıktılarını yorumlayarak, gerçek dünyada nasıl kullanılabileceğini ve hangi koşullarda güncellenmesi gerektiğini açıklamak amaçlanmıştır.
Model yalnızca teknik olarak değil, iş kararlarına katkı sağlayacak biçimde değerlendirilmiştir.

Uygulama Alanı ve Kullanım Önerileri:

Uygulama Alanı	Açıklama
Meteoroloji / Hava Tahmini	Günlük "yağış olacak mı?" tahminlerinde destek aracı olarak kullanılabilir.
Tarım Planlaması	Yağmur olasılığına göre sulama, ekim ve hasat planları yapılabilir.
Enerji Sektörü	Güneşli / yağmurlu gün tahmini, yenilenebilir enerji üretim planlarını optimize edebilir.
Ulaşım ve Lojistik	Yağış riskine göre trafik yoğunluğu, rota planlama ve teslimat zamanlaması geliştirilebilir.

Modelin Güncellenmesi (Model Maintenance)

Makine öğrenmesi modelleri, çevresel veriler değiştikçe yeniden eğitilmelidir.
Bu nedenle aşağıdaki strateji önerilir:

- Periyodik Güncelleme: Model, her 6 ayda bir yeni meteorolojik verilerle yeniden eğitilmeli.
- Performans Takibi: Accuracy değeri %80'in altına düşerse yeniden eğitilmesi gerekir.
- Otomatik Pipeline: KNIME üzerinde AutoML veya Python scheduler ile düzenli yeniden eğitim yapılabilir.