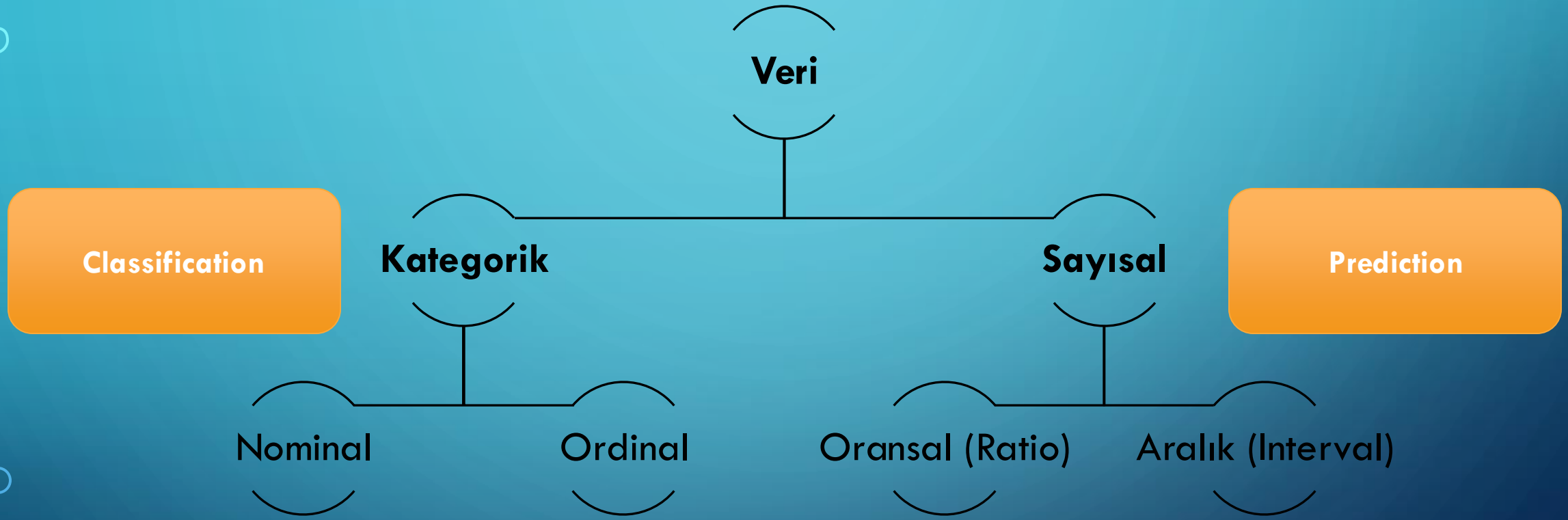
VERİ ÖN İŞLEME

ECEHAN ÜZGÜN


Veri
Classification
Kategorik
Sayısal
Prediction
Nominal
Ordinal
Oransal (Ratio)
Aralık (Interval)

VERİ BİRLEŞTİRME

1. **Dosya okuma:** CSV dosyasından verileri Pandas DataFrame formatında içe aktarıyoruz.
2. **Eksik veri doldurma:** SimpleImputer kullanarak eksik değerleri sütunların ortalamasıyla dolduruyoruz.
3. **Kategorik verileri sayısala çevirme:** LabelEncoder ve OneHotEncoder kullanarak ülkeleri sayısal forma dönüştürüyoruz.
4. **Yeni DataFrame'ler oluşturma:** Kategorik ve sayısal verileri ayrı DataFrame'lere koyuyoruz.
5. **DataFrame'leri birleştirme:** pd.concat() ile sütun bazında birleştirerek tam bir veri seti oluşturuyoruz.

VERİ KÜMESİNİN EĞİTİM VE TEST OLARAK BÖLÜNMESİ

```
from sklearn.model_selection import train_test_split

x_train, x_test, y_train, y_test = train_test_split(s,sonuc3,test_size=0.33,
random_state=0)
```

X: Bağımsız değişkenler → s
Y: Bağımlı değişkenler → sonuc3
Test_size=0.33 : Verinin %33'ü test verisi, %67'si eğitim verisi olacak.
Random_state = 0 : Aynı verinin tekrar kullanılması için rastgeleliği kontrol eden bir parametre.

VERİYİ EĞİTİM VE TEST SETİNE AYIRMA AMACIMIZ NEDİR?

Makine öğrenimi modellerini oluştururken, modelin gerçek dünyada yeni verilerle nasıl performans göstereceğini test etmek için elimizdeki verileri **eğitim (training)** ve **test (testing)** setleri olarak ikiye ayırırız.

Amaç: Modelin sadece eğitildiği veriyi ezberlemesini (overfitting) önlemek ve yeni veriler üzerinde nasıl çalışacağını değerlendirmektir.

GERÇEK PROJELERDE NEDEN KULLANILIR?

1. **Modelin Performansını Değerlendirme**

❖ Model eğitim setiyle eğitilir.

❖ Daha sonra test setindeki veriler modele verilerek doğru tahmin yapıp yapmadığı kontrol edilir.

❖ Böylece, modelin genelleme (generalization) yapabilme yeteneğini ölçeriz.

📌 Gerçek Dünya Senaryosu:

Bir kredi risk değerlendirme modelini düşünelim.

- Eğitim seti: Bankanın geçmiş müşteri verileri (Kredi çekmiş ve ödemiş/ödeyememiş müşteriler)
- Test seti: Yeni gelen müşterilerin verileriyle modelin doğruluğunu test ederiz.
- Eğer model test setinde düşük başarı gösteriyorsa, model gerçek dünyaya da kötü tahminler yapacaktır.

2. EZBERLEME (OVERFITTING) VE EKSİK ÖĞRENMEYİ (UNDERFITTING) ÖNLEME

Overfitting: Modelin sadece eğitim setindeki verileri ezberleyip, yeni veriler üzerinde başarısız olmasıdır.

- Test setiyle kontrol etmezsek model eğitildiği verilere aşırı uyum sağlayabilir ve yeni verilerde kötü sonuçlar verebilir.

📌 Gerçek Dünya Senaryosu:

- Bir spam e-posta tespit modeli geliştirdiğimizi düşünelim.
- Eğer sadece belirli kelimeler (örn. para kazan, hediye, indirim) üzerine eğitirsek, yeni gelen spam e-postaları tespit edemeyebiliriz.
- Bu yüzden, modeli test verisiyle denememiz gerekir.

3. MODELİN GERÇEK VERİLERE HAZIRLANMASINI SAĞLAMA

- Eğitim seti, modelin öğrendiği verileri içerirken; test seti gerçek dünyadaki bilinmeyen verileri temsil eder.
- Model test verisi üzerinde başarılı olursa, gerçek dünyadaki veriler üzerinde de başarılı olma ihtimali yüksektir.
- Modelin hata oranlarını ölçmek için bu işlem yapılır.

📌 Gerçek Dünya Senaryosu:

- Bir hastalık teşhis modeli düşünelim.
- Eğer modeli sadece belirli bir hastane verisiyle eğitirsek, farklı hastalardan gelen verileri tanımayabilir.
- Test verisini farklı hastanelerden alarak modelin gerçekten genelleyip genellemediğini kontrol etmeliyiz.

TEST VERİSİ İLE EĞİTİM VERİSİ ARASINDAKİ FARK

Özellik	Eğitim Verisi (Training Data)	Test Verisi (Testing Data)
Kullanım Amacı	Modelin öğrenmesi için kullanılır.	Modelin doğruluğunu test etmek için kullanılır.
Modelin Gördüğü Veriler	Model bu verileri öğrenir.	Model bu verileri daha önce görmemiştir.
Etkisi	Modelin parametrelerini ayarlamak için kullanılır.	Modelin genelleme yeteneğini ölçmek için kullanılır.

```
from sklearn.model_selection import train_test_split

x_train, x_test, y_train, y_test = train_test_split(s,sonuc3,test_size=0.33,
random_state=0)
```

- %33 Test oranı seçilmesi, modelin eğitim için yeterli veriye sahip olması ve test için de yeterli gözlem içermesi içindir.
- Daha büyük veri setlerinde %10 veya %20 test verisi yeterli olabilir.

SONUÇ

Makine öğrenimi modelinin genel performansını ölçmek için veriyi eğitim ve test setlerine ayırırız.

Overfitting (ezberleme) ve underfitting (eksik öğrenme) riskini azaltırız.

Modelin bilinmeyen verilerde başarılı olup olmadığını değerlendiririz.

Gerçek dünyada modelin iyi çalışmasını sağlarız.

Gerçek bir projede modelin test setinde kötü sonuçlar vermesi, modelin yeniden eğitilmesi gerektiğini gösterir.