

İstatistik

IBM Modeller ve SPSS

IBM SPSS (Statistical Package for the Social Sciences)

- **SPSS** özellikle sosyal bilimler, pazar araştırmaları, sağlık, eğitim gibi alanlarda **istatistiksel analiz** yapmak için kullanılan bir yazılımdır.
- IBM, SPSS'i 2009 yılında satın aldı; bu yüzden bugün resmi ismi "IBM SPSS Statistics"tir.
- SPSS ile:
 - Veri analizi (tanımlayıcı istatistikler, regresyon analizleri, faktör analizleri vb.)
 - Veri madenciliği
 - Anket analizi
 - Hipotez testi
 - Grafik ve tablo oluşturma yapılabilir.
- Kullanımı, sürükle-bırak veya menü tabanlıdır, ayrıca syntax (kod) yazılarak da çalışılabilir.

IBM Modeller (IBM SPSS Modeler)

- **IBM SPSS Modeler**, daha çok **veri madenciliği** ve **makine öğrenimi** modelleri geliştirmek için kullanılan bir araçtır.
- Analistlerin ve veri bilimcilerin görsel bir arayüz üzerinden **makine öğrenmesi** ve **yapay zeka** modelleri kurmasını sağlar.
- SPSS Modeler ile:
 - Karar ağaçları, regresyonlar, sinir ağları, kümeleme gibi **gelişmiş modelleme teknikleri**
 - Büyük veri (Big Data) analizi
 - Otomatik makine öğrenimi (AutoML)
 - Metin analitiği
 - Zaman serisi analizi yapılabilir.
- Özellikle finans, sağlık, telekomünikasyon gibi alanlarda tahminleme ve müşteri analizi için kullanılır.

Özellik	IBM SPSS Statistics	IBM SPSS Modeler
Ana kullanım amacı	İstatistiksel analiz ve raporlama	Veri madenciliği ve makine öğrenmesi
Kullanıcı arayüzü	Menü tabanlı ve kod (syntax) destekli	Görsel akış (flow-based) modelleme
Analiz türü	Tanımlayıcı istatistikler, hipotez testleri, regresyon analizleri	Tahmine dayalı modeller, sınıflandırma, kümeleme
Programlama gerekliliği	Yok (isteğe bağlı syntax yazılabilir)	Yok (görsel sürükle-bırak yöntemi)
Veri kaynakları	Genellikle küçük/orta ölçekli veri setleri	Büyük veri kaynakları, veri ambarları, Hadoop, NoSQL
Makine öğrenmesi desteği	Sınırlı (bazı regresyon ve karar ağacı)	Geniş (sinir ağları, random forest, SVM, vb.)
Hedef kullanıcılar	Sosyal bilimciler, akademisyenler, analistler	Veri bilimciler, analitik uzmanlar
Örnek kullanım alanı	Anket analizi, sosyal araştırmalar	Müşteri segmentasyonu, churn tahmini
Ekstra özellikler	Raporlama, veri temizleme araçları	Otomatik model seçimi (AutoML), metin madenciliği
Çıktılar	Tablolar, grafikler, özet raporlar	Model akışları, tahmin skorları, görselleştirme

Örnekler

1. IBM SPSS Statistics Uygulaması (Basit İstatistiksel Analiz)

Senaryo:

Bir öğretmen, öğrencilerinin sınav notlarını analiz etmek istiyor. Ortalama notu, standart sapmayı ve kaç öğrencinin geçtiğini görmek istiyor.

Veri Örneği:

(Excel dosyası gibi bir veri seti düşün.)

Öğrenci	Sınav Notu
A	78
B	85
C	62
D	90
E	55

SPSS'te Yapılacaklar:

1. SPSS'i aç.
2. **Variable View** kısmında:
 - **Öğrenci** (string), **Sınav_Notu** (numeric) değişkenlerini tanımla.
3. **Data View** kısmında bu verileri gir.
4. Menüden:
Analyze > Descriptive Statistics > Descriptives
seç ve **Sınav_Notu** değişkenini seç.
5. Ortalama, standart sapma, min-maks değerlerini görebilirsin.
6. Ayrıca:
Analyze > Descriptive Statistics > Frequencies
ile belirli bir baraj (örneğin 60) üstünde kaç kişi olduğunu görebilirsin.

2. IBM SPSS Modeler Uygulaması (Basit Tahmin Modeli)

Senaryo:

Bir banka, müşterilerinin kredi başvurularının onaylanıp onaylanmayacağını tahmin etmek istiyor.

Veri Örneği:

Yaş	Gelir	Kredi Skoru	Onay Durumu (Evet/Hayır)
25	3000	650	Hayır
45	8000	720	Evet
35	5000	680	Evet
28	4000	600	Hayır

Modeler'da Yapılacaklar:

1. IBM SPSS Modeler'ı aç.
2. **Read** düğümünden veri setini yükle (CSV veya Excel dosyasından).
3. **Type** düğümünde:
 - "Onay Durumu" hedef (target) değişken olarak işaretlenir.
4. **Model** sekmesinden bir karar ağacı algoritması (örneğin C5.0) seç.
5. Modeli oluştur ve çalıştır (build and run).
6. Çıkan ağaç yapısında:
 - Gelir > 6000 TL üstüyse ve kredi skoru > 700 ise, **kredi onaylanıyor** gibi kurallar görebilirsin.
7. **Evaluation** düğümü ile modelin doğruluk oranını test edebilirsin.

Veri Ön İşleme (Data Preprocessing)

Bu süreç, hem **IBM SPSS Modeler** hem de genel veri bilimi projelerinde aynıdır.

Adım	Anlamı	Neden Yapılır?
Outlier Operations	Aykırı değerleri (normalden çok sapmış uç değerler) tespit edip düzeltme veya çıkarma.	Modelin sapmasını ve yanlış tahminler yapmasını engeller.
Missing Value Operations	Eksik verileri doldurma (imputation) veya eksik satırları atma.	Modelin hata almadan çalışması için eksiklikler çözülür.
Transformation	Verileri log alma, root alma gibi işlemlerle dönüştürmek.	Dağılımı normalleştirmek veya değişkenler arası ilişkiyi güçlendirmek için.
Standardisation	Verileri ortalama=0, standart sapma=1 olacak şekilde ölçeklemek.	Özellikle mesafe tabanlı algoritmalar (k-means, kNN gibi) için çok kritiktir.
Feature Extracting	Ham veriden yeni anlamlı değişkenler (özellikler) üretmek.	Modelin öğrenebileceği bilgiyi zenginleştirir.
Feature Importance	Hangi değişkenlerin daha önemli olduğunu analiz etmek.	Önemsiz değişkenleri eleyerek model performansı artırılır.



Kısaca:

Bu adımlar, **güçlü ve sağlam bir makine öğrenmesi modeli** kurmanın temelidir.

Eğer bu işlemleri doğru yaparsan, kurduğun model çok daha **doğru tahminler yapar** ve **genelleme yeteneği yüksek olur**.

📈 SPSS Modeler'da Veri Ön İşleme ve Modelleme Akışı

İşte genel şemamız:

[Read File]
↓
[Outlier Analysis]
↓
[Missing Value Analysis]
↓
[Data Transformation]
↓
[Standardisation]
↓
[Feature Extraction]
↓
[Feature Selection (Importance)]
↓
[Modeling (Decision Tree, Logistic Regression, etc.)]
↓
[Evaluation]

🎯 Akışta Her Düğüm Ne Yapar?

Düğüm (Node)	Açıklaması
Read File	Veriyi (CSV, Excel, SQL vb.) içeri aktarır.
Outlier Analysis	Uç (aykırı) değerleri tespit eder. (İstersen bunları silebilir ya da düzeltebilirsin.)

Missing Value Analysis	Eksik değerleri analiz eder, doldurur veya kayıtları atar.
Data Transformation	Gerekirse veriyi dönüştürür. (Örn: log, karekök gibi.)
Standardisation	Verileri standart ölçeğe getirir. (Özellikle kNN, k-Means gibi algoritmalar için.)
Feature Extraction	Yeni değişkenler yaratır. (Örn: tarih verisinden ay bilgisi çıkarma gibi.)
Feature Selection	Hangi değişkenlerin en önemli olduğunu belirler. Gereksiz olanları eleyebiliriz.
Modeling	Bir makine öğrenmesi modeli kurarız. (Örn: Karar Ağacı, Regresyon, Destek Vektör Makineleri vb.)
Evaluation	Modelin doğruluk oranını ölçeriz. (Confusion Matrix, ROC Curve gibi analizler yaparız.)

SPSS Modeller'da Bunu Yaparken Kullanacağın Düğümler

Modeller menüsünde her adım için ayrı bir düğüm vardır. Mesela:

- **Record Ops** → Outlier Analysis, Sample, Partition
- **Field Ops** → Missing Values, Derive, Reclassify
- **Transform** → Binning, Normalize, Standardize
- **Model** → Decision Tree (C&R Tree, CHAID, C5.0), Neural Net, Regression
- **Evaluation** → Analysis Node, Evaluation Node

Bu düğümleri sürükleyip aralarına bağlantı çizmen yeterli.

Modeller'da çalışma prensibi Lego gibi: Düğümleri ekle, bağla, çalıştır!

Basit Senaryo ile Açıklama (Örnek)

Diyelim ki bir banka verisi var:

Müşteri_ID	Yaş	Gelir	Kredi Skoru	Borç Durumu	Onaylandı (Hedef)
1	25	3000	620	5000	Hayır
2	45	8000	720	2000	Evet
...

Senin akışın şunları yapacak:

- Yaşı 150 olan bir aykırı müşteri varsa → **Sil**.
- Gelir bilgisi eksikse → **Ortalama ile doldur**.
- Gelir verisini → **Log dönüşümüne sok**.
- Tüm sayısal verileri → **Standardize et**.
- Gelir / Borç oranı gibi bir yeni değişken yarat → **(Feature Extraction)**.
- En önemli değişkenleri belirle (örneğin Kredi Skoru en önemlisi çıktı) → **Feature Importance**.
- Sonra **Karar Ağacı (C5.0)** ile kimin kredisi onaylanır onu tahmin et.
- Son adımda modeli değerlendirip doğruluk oranına bak.