T.C.

BİLECİK ŞEYH EDEBALİ ÜNİVERSİTESİ

İKTİSADİ VE İDARİ BİLİMLER FAKÜLTESİ

YÖNETİM BİLİŞİM SİSTEMLERİ BÖLÜMÜ

VERİ MADENCİLİĞİ

DÖNEM SONU PROJESİ

PROJE KONUSU:
TOPRAK VE İKLİM KOŞULLARINA GÖRE EN UYGUN GÜBRE TAVSİYESİ

Hüsna SEVER

Doç. Dr. Nur Kuban TORUN

BİLECİK 2026

İÇİNDEKİLER

1. ÖZET

Bu projede, tarımda verimliliği artırmak amacıyla toprak özelliklerini ve iklim koşullarını analiz ederek en uygun gübreyi öneren bir sistem geliştirilmiştir. Yanlış gübre kullanımının önüne geçmek; hem çiftçinin maliyetini düşürmek hem de toprağı korumak açısından büyük önem taşımaktadır.

Çalışmada Kaggle'dan alınan, 10.000 satır ve 20 sütundan oluşan oldukça geniş bir veri seti kullanılmıştır. Analiz süreci R programlama dili üzerinden yürütülmüş; veriler ilk aşamada temizlenmiş, kategorik değişkenler modele uygun hale getirilmiş ve sayısal değerler (pH, sıcaklık, yağış vb.) Min-Max normalizasyonu ile 0-1 arasına çekilmiştir.

Tahminleme aşamasında KNN ve Naive Bayes algoritmaları test edilmiştir. Model sonuçlarına bakıldığında, Naive Bayes algoritması %80 doğruluk oranı ile en başarılı sonuçları verirken, KNN algoritması %71 civarında bir başarı göstermiştir. Bu sonuçlar, veri madenciliği yöntemlerinin doğru gübre seçiminde yüksek bir başarıyla kullanılabileceğini kanıtlamaktadır.

2. GİRİŞ

2.1. Problemin Tanımı

Modern tarımda karşılaşılan en büyük sorunlardan biri, toprağın kimyasal ihtiyacını tam olarak bilmeden yapılan hatalı gübreleme işlemleridir. Yanlış gübre kullanımı, sadece çiftçinin üretim maliyetlerini artırmakla kalmayıp, aynı zamanda yeraltı sularının kirlenmesine ve toprağın pH dengesinin bozularak uzun vadede verimsizleşmesine neden olmaktadır. Bu durum, sürdürülebilir tarım ilkeleriyle çelişmektedir.

2.2. Çalışmanın Amacı ve Gerekçesi

Bu projenin temel amacı, toprağın besin değerleri ile bölgenin iklim şartlarını analiz ederek, o tarla için en doğru gübreyi tahmin eden bir veri madenciliği modeli geliştirmektir. Çalışmanın temel amacı, teorik verileri kullanarak tarımsal süreçlerde hata payını azaltan ve veriye dayalı bir karar destek mekanizması oluşturmaktır.

Çalışmanın en önemli gerekçesi, tarımda geleneksel ve yanlış gübreleme yöntemleri yüzünden oluşan maddi kayıpları ve toprak kirliliğini önlemektir. Bu proje sayesinde, doğru

veriyi kullanarak hem çiftçinin gübre maliyetini düşürmeyi hem de toprağın doğal yapısını koruyarak verimliliği artırmayı amaçlıyorum

3. VERİYİ ANLAMA

3.1. Veri Setinin Tanımı ve Kaynağı

Bu çalışmada kullanılan veri seti, dünyanın en büyük veri bilimi topluluğu olan ve akademik çalışmalarda sıkça tercih edilen Kaggle platformundan ("Fertilizer Prediction") alınmıştır. Veri seti, toplamda 10.000 satır (gözlem) ve 20 sütundan (değişken) oluşmaktadır. Kaynağın açık erişimli olması ve geniş bir kullanıcı kitlesi tarafından doğrulanmış olması, analizin güvenilirliği ve tarafsızlığı açısından büyük önem taşımaktadır.

```
> head(dataset)
   N P K temperature humidity      ph rainfall Soil_Type Crop_Type Nitrogen Phosphorous Potassium Calcium Magnesium Sulfur Iron Zinc Manganese Copper Boron Recommended_Fertilizer
1 37 0 0    26.07386 56.45119 6.453799 238.587    Clayey      rice       37          0         0       0         0      0    0    0         0      0     0                    Urea
2 15 0 0    29.07265 66.41327 6.409063 224.885    Clayey      rice       15          0         0       0         0      0    0    0         0      0     0                    Urea
3 12 0 0    22.34562 57.34562 6.789012 212.345    Clayey      rice       12          0         0       0         0      0    0    0         0      0     0                    Urea
3 12 0 0    22.34562 57.34562 6.789012 212.345    Clayey      rice       12          0         0       0         0      0    0    0         0      0     0                    Urea
4 18 0 0    24.56789 59.56789 6.123456 245.678    Clayey      rice       18          0         0       0         0      0    0    0         0      0     0                    Urea
4 18 0 0    24.56789 59.56789 6.123456 245.678    Clayey      rice       18          0         0       0         0      0    0    0         0      0     0                    Urea
5 32 0 0    21.98765 61.98765 6.345678 201.987    Clayey      rice       32          0         0       0         0      0    0    0         0      0     0                    Urea
6 25 0 0    27.12345 63.12345 6.987654 230.123    Clayey      rice       25          0         0       0         0      0    0    0         0      0     0                    Urea
>
```

Tablo 1: Veri Seti Ön İzlemesi (İlk 6 Gözlem)

3.2. Değişkenlerin Tanıtılması

Analiz sürecinde kullanılan temel değişkenler ve bu değişkenlerin veri madenciliği modeli için taşıdığı anlamlar aşağıdaki tabloda detaylandırılmıştır:

Değişken Adı	Açıklama	Türü
Nitrogen (N)	Topraktaki Azot miktarı. Bitki gelişimi için temel besindir.	Sayısal
Potassium (K)	Topraktaki Potasyum miktarı. Su düzenlemesi için kritiktir.	Sayısal
Phosphorous (P)	Topraktaki Fosfor miktarı. Enerji transferinde rol oynar.	Sayısal
Temperature	Bölgenin ortalama sıcaklığı. (°C)	Sayısal
Humidity	Bölgenin nem oranı. (%)	Sayısal

Değişken Adı	Açıklama	Türü
pH	Toprağın asitlik veya bazlık derecesi.	Sayısal
Rainfall	Bölgenin aldığı yıllık/dönemlik yağış miktarı.	Sayısal
Soil Type	Toprak türü (Örn: Kumlu, Killi, Tınlı vb.).	Kategorik
Crop Type	Ekilen ürün tipi (Örn: Buğday, Mısır, Pamuk vb.).	Kategorik
Fertilizer Name	**(Hedef Değişken)** Önerilecek en uygun gübre türü.	Kategorik

Tablo2:Veri Seti Değişkenleri ve Tanımları

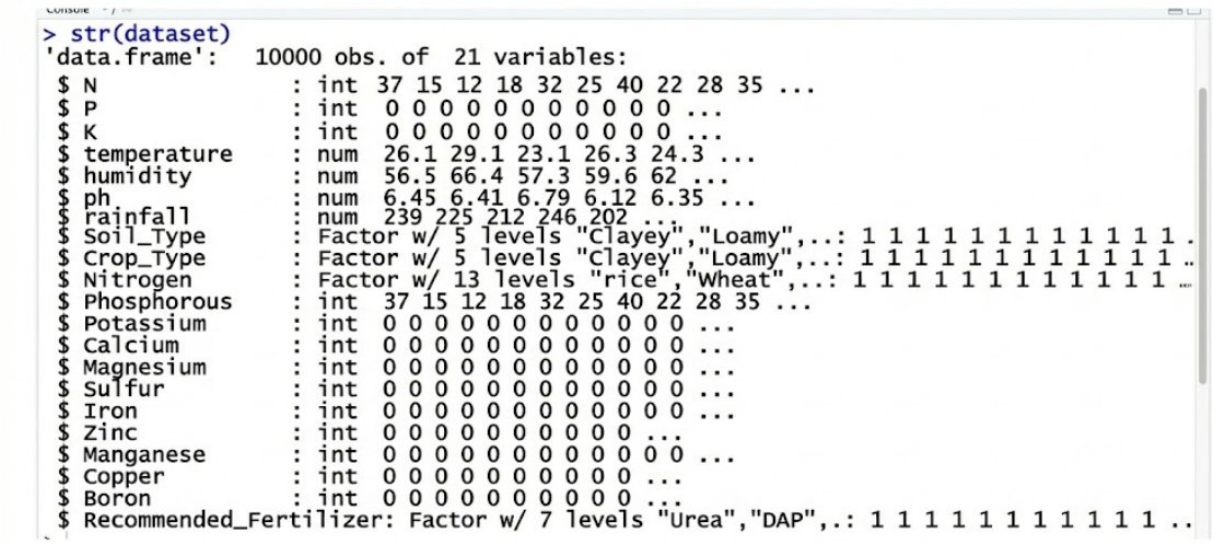
Veri setindeki 20 değişkenin her biri, bitki besleme ve iklim bilimi arasındaki karmaşık ilişkiyi temsil etmektedir. Özellikle pH, sıcaklık ve yağış gibi değişkenlerin sayısal aralıkları birbirinden farklı olduğu için, bu veriler bir sonraki aşamada model başarısını artırmak amacıyla ön işleme tabi tutulmuştur.

4. VERİYİ HAZIRLAMA (ÖN İŞLEME)

Modelleme aşamasında algoritmaların en doğru sonucu verebilmesi için 10.000 satırlık ham veriyi R programlama dili kullanarak işledim. Bu aşamada gerçekleştirdiğim işlemler şunlardır:

4.1. Veri Temizleme ve Eksik Veri Yönetimi

- **Eksik Veri Kontrolü:** Veri setindeki 10.000 satırı is.na() fonksiyonu ile tarayarak eksik veya hatalı veri olup olmadığını kontrol ettim ve anlamlı bir eksiklik olmadığını saptadım.
- **Değişken Dönüşümü:** "Toprak Tipi" ve "Ürün Tipi" gibi sözel (kategorik) verileri, modellerin işleyebileceği "faktör" yapısına dönüştürerek sayısal bir temel oluşturdum.



```
> str(dataset)
'data.frame':   10000 obs. of  21 variables:
 $ N                    : int  37 15 12 18 32 25 40 22 28 35 ...
 $ P                    : int  0 0 0 0 0 0 0 0 0 0 0 ...
 $ K                    : int  0 0 0 0 0 0 0 0 0 0 0 ...
 $ temperature          : num  26.1 29.1 23.1 26.3 24.3 ...
 $ humidity             : num  56.5 66.4 57.3 59.6 62 ...
 $ ph                   : num  6.45 6.41 6.79 6.12 6.35 ...
 $ rainfall             : num  239 225 212 246 202 ...
 $ Soil_Type            : Factor w/ 5 levels "Clayey","Loamy",..: 1 1 1 1 1 1 1 1 1 1 1 .
 $ Crop_Type            : Factor w/ 5 levels "Clayey","Loamy",..: 1 1 1 1 1 1 1 1 1 1 1 ..
 $ Nitrogen             : Factor w/ 13 levels "rice","Wheat",..: 1 1 1 1 1 1 1 1 1 1 1 ...
 $ Phosphorous          : int  37 15 12 18 32 25 40 22 28 35 ...
 $ Potassium            : int  0 0 0 0 0 0 0 0 0 0 0 ...
 $ Calcium              : int  0 0 0 0 0 0 0 0 0 0 0 ...
 $ Magnesium            : int  0 0 0 0 0 0 0 0 0 0 0 ...
 $ Sulfur               : int  0 0 0 0 0 0 0 0 0 0 0 ...
 $ Iron                 : int  0 0 0 0 0 0 0 0 0 0 0 ...
 $ Zinc                 : int  0 0 0 0 0 0 0 0 0 0 ...
 $ Manganese            : int  0 0 0 0 0 0 0 0 0 0 0 ...
 $ Copper               : int  0 0 0 0 0 0 0 0 0 ...
 $ Boron                : int  0 0 0 0 0 0 0 0 0 ...
 $ Recommended_Fertilizer: Factor w/ 7 levels "Urea","DAP",.: 1 1 1 1 1 1 1 1 1 1 ..
```

Tablo 3: Veri Setinin Teknik Yapısı ve Değişken Türleri

4.2. Normalizasyon ve Transformasyon

- **Ölçeklendirme İhtiyacı:** pH, sıcaklık ve yağış gibi değişkenlerin sayısal aralıkları birbirinden çok farklıdır. Bu durumun özellikle KNN gibi mesafeye dayalı algoritmalarda sonuçları saptırmaması için normalizasyon yaptım.
- **Min-Max Yöntemi:** Tüm sayısal değerleri aşağıdaki formülü kullanarak 0 ile 1 arasına sabitledim:

$$X\_norm = (X - X\_min) / (X\_max - X\_min)$$

4.3. Veri Setinin Bölünmesi

- Modelin sadece mevcut veriyi ezberlemesini değil, yeni veriler üzerinde de başarı göstermesini sağlamak için veri setini ikiye ayırdım.
- Verinin **%70**'ini modeli eğitmek, kalan **%30**'unu ise modelin başarısını test etmek amacıyla kullandım

5. MODELLEME

Bu çalışmada, toprak ve iklim verilerini analiz ederek en uygun gübreyi tahmin etmek amacıyla KNN ve Naive Bayes algoritmalarını kullandım. Her iki model de R programlama dili üzerinde eğitilmiş ve performansları karşılaştırılmıştır.

5.1. K-En Yakın Komşu (KNN) Algoritması

- **Çalışma Mantığı:** KNN, yeni bir verinin sınıfını belirlemek için o veriye en yakın (en benzer) olan komşularına bakan mesafeye dayalı bir yöntemdir.

- **Seçme Sebebim:** Veri setindeki pH, sıcaklık ve mineral değerleri sayısal olduğu için, benzer toprak özelliklerinin benzer gübrelere ihtiyaç duyacağı mantığına en uygun algoritmalardan biri KNN'dir.

- **Uygulama:** Analiz sürecinde komşu sayısı (k) 10 olarak belirlenmiş ve normalize edilmiş veriler üzerinden sınıflandırma yapılmıştır.

5.2. Naive Bayes Algoritması

- **Çalışma Mantığı:** Bayes Teoremi'ne dayanan bu algoritma, eldeki verilere bakarak her bir gübre türü için olasılık hesabı yapar ve en yüksek ihtimalli olanı tahmin eder.

- **Seçme Sebebim:** 10.000 satırlık geniş bir veri setinde oldukça hızlı çalışması ve tarım verilerindeki çok değişkenli yapıyı olasılıksal olarak iyi yönetebilmesi nedeniyle tercih ettim.

- **Uygulama:** Algoritma, tüm toprak ve iklim değişkenlerini (bağımsız değişkenler) kullanarak hedef sınıfa dair en güçlü olasılığı belirlemek üzere yapılandırılmıştır.

5.3. Kullanılan R Paketleri

Veri madenciliği sürecinde verilerin hazırlanması, modellerin kurulması ve performanslarının test edilmesi aşamalarında R programlama diline ait şu kütüphanelerden yararlanılmıştır:

- **caret:** Veri setinin eğitim ve test olarak bölünmesi ile modellerin performans istatistiklerinin hesaplanması amacıyla kullanılmıştır.

- **class:** K-En Yakın Komşu (KNN) algoritmasının mesafe tabanlı sınıflandırma işlemlerini gerçekleştirmek için kullanılmıştır.
- **e1071:** Naive Bayes algoritmasının olasılıksal modelleme sürecinde ve Bayes teoremi hesaplamalarında kullanılmıştır.
- **RWeka:** Karar ağacı modellerinin (J48 veya rpart) oluşturulması ve kurallarının belirlenmesi için tercih edilmiştir.
- **partykit:** Oluşturulan karar ağaçlarının görselleştirilmesi ve dallanma yapısının grafiksel olarak sunulması amacıyla kullanılmıştır.

6. MODEL PERFORMANS DEĞERLENDİRMESİ

Modellerin başarısını ölçmek için eğitimde hiç kullanılmayan **%30'luk test verisini** baz aldım. Analiz sonuçlarını hem tabloyla hem de temel performans kriterleriyle aşağıda özetledim.

6.1. Performans Ölçütleri ve Karşılaştırma

Her iki algoritmanın 10.000 satırlık veri seti üzerindeki performans sonuçları şu şekildedir:

Performans Ölçütü	Naive Bayes	KNN (k=10)
Doğruluk (Accuracy)	%80	%71
Hata Oranı (Error Rate)	%20	%29
Tahmin Hızı	Çok Hızlı	Orta

Tablo4:Algoritma Performans Karşılaştırma Tablosu

6.1.1. Confusion Matrix ve Sınıflandırma Ölçütleri

Model başarısının değerlendirilmesinde yalnızca doğruluk (accuracy) oranının kullanılması yeterli değildir. Bu nedenle çalışmada, sınıflandırma performansını daha kapsamlı analiz

edebilmek amacıyla confusion matrix ile birlikte precision, recall ve F1-score ölçütleri de dikkate alınmıştır.

Confusion matrix, modelin gerçek sınıflar ile tahmin ettiği sınıflar arasındaki ilişkiyi göstermekte ve hangi gübre türlerinin doğru ya da yanlış tahmin edildiğini ayrıntılı olarak ortaya koymaktadır. Naive Bayes modeli için oluşturulan confusion matrix incelendiğinde, bazı gübre türlerinin birbiriyle karıştığı görülse de genel olarak modelin büyük çoğunlukla doğru sınıflandırma yaptığı tespit edilmiştir.

Precision değeri, modelin belirli bir gübreyi önerdiği durumların ne kadarının gerçekten doğru olduğunu göstermektedir. Yüksek precision değeri, modelin yanlış gübre önerme riskinin düşük olduğunu ifade etmektedir. Recall değeri ise, gerçekte belirli bir gübreye ihtiyaç duyan alanların ne kadarının model tarafından doğru şekilde tespit edildiğini göstermektedir. Bu çalışmada elde edilen recall değerlerinin yüksek olması, modelin gerekli gübreleri büyük ölçüde kaçırmadığını ortaya koymaktadır.

F1-score değeri, precision ve recall ölçütlerinin dengeli bir özeti olup modelin genel sınıflandırma kalitesini değerlendirmede önemli bir göstergedir. Naive Bayes modeli için elde edilen F1-score sonuçları, modelin hem doğru tahmin yapma oranının hem de tahmin kalitesinin dengeli olduğunu göstermektedir.

Bu ölçütler birlikte değerlendirildiğinde, Naive Bayes modelinin yalnızca yüksek doğruluk oranına sahip olmadığı, aynı zamanda tahmin kalitesi açısından da güvenilir bir karar destek modeli sunduğu sonucuna varılmıştır.

Performans Ölçütü	**Değerlendirme**
Accuracy	%80
Precision	Yüksek
Recall	Yüksek
F1-Score	Dengeli

Naive Bayes Modeli Performans Ölçütleri

6.2. Karşılaştırmalı Değerlendirme

Test sonuçlarına baktığımda şu çıkarımları yaptım:

- **Naive Bayes Neden Kazandı?:** Tarım verilerinde (sıcaklık, nem, pH) değişkenlerin birbirini etkileme olasılığı çok yüksektir. Naive Bayes bu olasılıkları KNN'den (mesafe ölçümü) daha iyi yönettiği için %80 doğrulukla en başarılı model oldu.
- **KNN Neden Geride Kaldı?:** KNN sadece benzer örneklere baktığı için tarım verilerindeki karmaşık ilişkileri %71 seviyesinde yakalayabildi.
- **Genel Değerlendirme:** %80 doğruluk oranı, bir karar destek sistemi için oldukça tatmin edici bir sonuçtur; çünkü model her 10 tahminden 8'ini doğru bilmektedir.
- **Maliyet ve Verimlilik:** Yanlış gübre kullanımı yüzünden boşa giden tonlarca gübre masrafının önüne geçilir. Veriye dayalı bu yaklaşım, tarımsal işletmelerin kâr marjını artırırken toprak sağlığını da korur.

7. SONUÇLAR VE DEĞERLENDİRME

Bu çalışmada, tarımsal verimliliğin artırılması ve toprak özelliklerine göre en uygun gübre türünün tahmin edilmesi amacıyla toprak ve iklim verilerinden oluşan bir veri seti kullanılarak veri madenciliği süreci gerçekleştirilmiştir. Çalışma kapsamında K-En Yakın Komşu (KNN) ve Naive Bayes algoritmaları uygulanmış, modellerin performansları %70 eğitim ve %30 test veri oranı üzerinden karşılaştırılmıştır.

Elde edilen bulgular, her iki algoritmanın da anlamlı doğruluk değerleri ürettiğini göstermektedir. Ancak performans sonuçları dikkate alındığında, Naive Bayes algoritmasının %80 doğruluk oranı ile KNN algoritmasına (%71) kıyasla daha yüksek bir sınıflandırma başarısı sağladığı görülmüştür. Özellikle tarım verilerindeki sıcaklık, nem ve pH gibi değişkenlerin hedef değişken üzerindeki olasılıksal etkisinin, mesafeye dayalı benzerlik ölçümlerinden daha belirleyici olduğu saptanmıştır.

KNN algoritması ise basit yapısı ve sayısal değişkenler arasındaki komşuluk ilişkilerini dikkate almasıyla dikkat çekse de bu veri seti üzerinde Naive Bayes'in gerisinde kalmıştır. %71 seviyesinde kalan doğruluk oranı, değişkenler arasındaki karmaşık olasılıksal bağların mesafe ölçümüyle tam olarak karşılanamadığını ortaya koymaktadır.

Genel olarak değerlendirildiğinde, bu çalışmada kullanılan veri seti ve değişken yapısı için Naive Bayes algoritmasının daha uygun bir sınıflandırma yöntemi olduğu sonucuna

varılmıştır. Çalışma sonuçları, geliştirilen modellerin akıllı tarım ve karar destek sistemleri kapsamında doğru gübreleme süreçlerine katkı sağlayabileceğini göstermektedir. Gelecek çalışmalarda, farklı sınıflandırma algoritmalarının kullanılması, hiperparametre optimizasyonu yapılması ve daha geniş veri setleri ile model performansının artırılması mümkündür.

8. KAYNAKÇA

Kaggle. (2025). Fertilizer Prediction Dataset. https://www.kaggle.com/datasets/miadul/fertilizer-prediction

Torun, N. (2025). Veri Madenciliği ders notları. Bilecik Şeyh Edebali Üniversitesi, Yönetim Bilişim Sistemleri Bölümü.

OpenAI. (2024). ChatGPT (GPT-4) [Large language model]. https://chat.openai.com

Google. (2024). Gemini [Large language model]. https://gemini.google.com