T.C.
İSTANBUL BEYKENT ÜNİVERSİTESİ
LİSANSÜSTÜ EĞİTİM ENSTİTÜSÜ

BİLGİSAYAR MÜHENDİSLİĞİ ANABİLİM DALI
BİLGİSAYAR MÜHENDİSLİĞİ BİLİM DALI

MAKİNE ÖĞRENMESİ ALGORİTMALARI KULLANILARAK TARIM ALANINDA ÜRÜN SEÇİMİNİ OPTİMİZE ETMEYE YÖNELİK BİR ÇALIŞMA

Yüksek Lisans Tezi

Tezi Hazırlayan
Mahmoud ABTINI

İstanbul, 2025

T.C.
İSTANBUL BEYKENT ÜNİVERSİTESİ
LİSANSÜSTÜ EĞİTİM ENSTİTÜSÜ

BİLGİSAYAR MÜHENDİSLİĞİ ANABİLİM DALI
BİLGİSAYAR MÜHENDİSLİĞİ BİLİM DALI

MAKİNE ÖĞRENMESİ ALGORİTMALARI KULLANILARAK TARIM ALANINDA ÜRÜN SEÇİMİNİ OPTİMİZE ETMEYE YÖNELİK BİR ÇALIŞMA

Yüksek Lisans Tezi

Tezi Hazırlayan:
Mahmoud ABTINI

Öğrenci No
2320003065

ORCID ID
0000-0003-1221-0032

Danışman
Dr. Öğr. Üyesi Talat FİRLAR

İstanbul, 2025

YEMİN METNİ

Yüksek Lisans Tezi olarak sunduğum **“Makine Öğrenmesi Algoritmaları Kullanılarak Tarım Alanında Ürün Seçimini Optimize Etmeye Yönelik Bir Çalışma”** başlıklı bu çalışmanın, bilimsel ahlak ve geleneklere uygun şekilde tarafımdan yazıldığını, bu tezdeki bütün bilgileri akademik ve etik kurallar içinde elde ettiğimi, yararlandığım eserlerin tamamının kaynaklarda gösterildiğini ve çalışmamın içinde kullanıldıkları her yerde bunlara atıf yapıldığını, patent ve telif haklarını ihlal edici bir davranışımın olmadığını belirtir ve bunu onurumla doğrularım. 21/10/2025

Mahmoud ABTINI

T.C.
İSTANBUL BEYKENT ÜNİVERSİTESİ
LİSANSÜSTÜ EĞİTİM ENSTİTÜSÜ MÜDÜRLÜĞÜ
TEZLİ YÜKSEK LİSANS SINAV TUTANAĞI

21/10/2025

Enstitümüz ***Bilgisayar Mühendisliği*** Anabilim Dalı ***Bilgisayar Mühendisliği*** Programı yüksek lisans öğrencilerinden ***2320003065*** numaralı ***Mahmoud ABTINI*** nin *"İstanbul Beykent Üniversitesi Lisansüstü Eğitim – Öğretim Yönetmeliği"* nin ilgili maddesine göre hazırlayarak, Enstitümüze teslim ettiği ***"Makine Öğrenmesi Algoritmaları Kullanılarak Tarım Alanında Ürün Seçimini Optimize Etmeye Yönelik Bir Çalışma"*** konulu tezini, Yönetim Kurulumuzun 08/10/2025 tarih ve 2025/41 sayılı toplantısında seçilen ve On-Line toplanan biz jüri üyeleri huzurunda, İstanbul Beykent Üniversitesi Lisansüstü Eğitim ve Öğretim Yönetmeliğinin 29. maddesinin 3. fıkrası gereğince Zoom programı aracılığıyla on-line olarak aday tarafından savunulmuş ve sonuçta adayın tezi hakkında ***"OYBİRLİĞİ"*** ile ***"KABUL"*** kararı verilmiştir.

İşbu tutanak, 2 nüsha olarak hazırlanmış ve Enstitü Müdürlüğü'ne sunulmak üzere tarafımızdan düzenlenmiştir.

DANIŞMAN
Dr. Öğr. Üyesi Ta*** Fİ***
(İstanbul Beykent Üniversitesi)

ÜYE
Dr. Öğr. Üyesi Ke*** Gö*** NA***
(İstanbul Beykent Üniversitesi)

ÜYE
Doç. Dr. Em*** SE***
(İstinye Üniversitesi)

Adı ve Soyadı	:	Mahmoud ABTINI
Danışmanı	:	Dr. Öğr. Üyesi Talat FİRLAR
Derecesi ve Tarihi	:	Yüksek Lisans (Tezli), 2025
Alanı	:	Bilgisayar Mühendisliği
Anahtar Kelimeler	:	Makine Öğrenmesi, Hassas Tarım, Mahsul Önerisi, Sınıflandırma Algoritmaları.

ÖZ

MAKİNE ÖĞRENMESİ ALGORİTMALARI KULLANILARAK TARIM ALANINDA ÜRÜN SEÇİMİNİ OPTİMİZE ETMEYE YÖNELİK BİR ÇALIŞMA

Çiftçilerin geleneksel yöntemlere dayanarak araziye ekilecek en uygun ürünü seçmeleri, çeşitli zorluklar ve belirsizlikler içermektedir. Bu durum, tarımsal verimliliği olumsuz etkileyebilir ve çiftçilerin gelirlerini azaltabilir. Dolayısıyla, bu sorunun çözülmesi, hem çiftçilerin refahı hem de ülkenin tarımsal sürdürülebilirliği açısından büyük önem taşımaktadır. Bu çalışmada, Büyük Veri Analitiği ve Makine Öğrenimi yöntemlerinin entegrasyonu ile çiftçilerin doğru mahsul seçiminde karşılaştıkları zorlukların hafifletilmesi hedeflenmektedir. Hindistan Gıda ve Tarım Odası tarafından sağlanan veritabanı üzerinden toplanan veriler, bu araştırmanın temelini oluşturmaktadır. Kullanılacak veriseti, mahsullerin ayrıntıları, toprak bileşimi, mahsulün büyümesi için uygun hava koşulları, sıcaklık, toprağın pH değeri ve yağış miktarı gibi çeşitli özellikleri içermektedir. Bu özellikler, çiftçilerin doğru mahsul seçiminde kritik öneme sahip bilgiler sunmakta ve yanlış mahsul seçimini azaltarak tarımsal üretkenliği artırma potansiyeli taşımaktadır. Araştırmada, Naive Bayes, Gradient Boosting, Artificial Neural Network, Random Forest Classifier ve Voting Classifier gibi uygun makine öğrenmesi yöntemleri belirlenmiştir. Bu algoritmalar, verisetinin özelliklerine göre en iyi performansı gösterecek şekilde seçilmiştir. Modelin geliştirilmesi sürecinde, ön işleme adımları gerçekleştirilmiş ve verilerin temizlenmesi, normalleştirilmesi ve uygun özelliklerin seçilmesi gibi işlemler yapılmıştır. Model eğitimi tamamlandıktan sonra, performans metrikleri olan kesinlik, duyarlılık ve F1 puanı gibi kriterler kullanılarak kapsamlı bir analiz gerçekleştirilmiştir Analiz sonuçları, hem eğitim hem de test verilerinde Naive Bayes algoritmasının %99,54'lük hassasiyet oranıyla en iyi performansı gösterdiğini ortaya koymuştur. Bu sonuç, Naive Bayes algoritmasının, tarımsal verilerin sınıflandırılması ve doğru mahsul önerisi konusunda etkili bir araç olduğunu göstermektedir. Sonuç bölümünde, bu yöntemlerin başarıları ve tarımsal uygulamalardaki potansiyel etkileri tartışılmıştır. Ayrıca, gelecekteki araştırmalar için önerilerde bulunulmuş ve makine öğrenimi tekniklerinin tarım sektöründe daha geniş bir şekilde nasıl uygulanabileceği üzerinde durulmuştur.

Name and Surname	:	Mahmoud ABTINI
Supervisor	:	Asst. Prof. Dr. Talat FİRLAR
Degree and Date	:	Master (Thesis), 2025
Major	:	Computer Engineering
Keywords	:	Machine Learning, Precision Agriculture, Crop Recommendation, Classification Algorithms.

ABSTRACT

A STUDY ON OPTIMIZING CROP SELECTION IN AGRICULTURE USING MACHINE LEARNING ALGORITHMS

Nowadays, with the development of much faster computers and machine algorithm systems, the use of artificial intelligence in the healthcare sector has increased significantly. The use of these algorithms in the field of medicine has led to significant developments. Artificial intelligence techniques are frequently used in the diagnosis and treatment of heart diseases. In the thesis study, using machine learning, one of the sub-branches of artificial intelligence, a data set of heart disease was taken as a classification problem and a prediction was made on the probability of the patients in this data set to have heart disease. The obtained data set was trained with the sub-algorithms of the supervised learning system, a separate model was created with each classification algorithm, and then heart disease was predicted by comparing these created models with the test data set. KNIME Analytics Platform was used as the application. Among the classification algorithms, LPROP Multilayer Algorithm, K-Means Clustering, Support Vector Machines, Naive Bayes, Decision Tree, Logistic Regression, neural network algorithms and collective learning methods were used. When the accuracy rate of the test data set was evaluated according to the results obtained in the trials and tests carried out with the application, the highest accuracy rate was achieved with the Navi Bayes method with a rate of 81.48%. In the study, the error matrices of the predictions were also evaluated and shown in tables

İÇİNDEKİLER

Sayfa No.

5. PERFORMANS ÖLÇÜTLERİ

TABLOLAR LİSTESİ

Sayfa No.

GRAFİKLER LİSTESİ

Sayfa No.

KISALTMALAR

ANN : Artificial Neural Network (Yapay Sinir Ağı)

EDA : Exploratory Data Analysis (Keşifsel Veri Analizi)

GBM : Gradient Boosting Machine (Gradyan Artırma Makinesi)

Hum : Humidity (Nem)

K : Potassium (Potasyum)

K-En : K-Nearest Neighbor (K-En Yakın Komşu Algoritması)

K-En : K-Nearest Neighbor (K-En Yakın Komşu Algoritması)

N : Nitrogen (Azot)

Naive Bayes : Naive Bayes (Saf Bayes Sınıflandırıcısı)

P : Phosphorus Fosfor)

PCA : Principal Component Analysis (Temel Bileşenler Analizi)

PH : Potential of Hydrogen (Toprak Asitliği / pH)

Rainfall : Rainfall (Yağış Miktarı)

RF : Random Forest (Rastgele Orman)

SVM : Support Sector Machine (Destek Vektör Makineleri)

SÖZLÜK

Azot (N), Fosfor (P), Potasyum (K): Bitki büyümesi için gerekli temel besin maddeleridir.

Derin Öğrenme (Deep Learning): Çok katmanlı yapay sinir ağları kullanarak karmaşık veri yapılarını analiz eden makine öğrenmesi alt alanıdır.

Destek Vektör Makineleri (Support Vector Machine, SVM): Verileri en iyi ayıran karar sınırlarını (hiperdüzlemleri) bulan sınıflandırma yöntemidir.

EDA (Exploratory Data Analysis): Veriyi anlamak için özet istatistikler, görselleştirmeler ve keşifsel yöntemlerin kullanıldığı analiz sürecidir.

Gradyan Artırma (Gradient Boosting Machine, GBM): Zayıf tahmin modellerini ardışık şekilde eğiterek güçlü bir tahmin modeli oluşturan yöntemidir.

K-En Yakın Komşu (K-Nearest Neighbor, K-NN): Yeni veriyi, en yakın komşularının sınıfına göre sınıflandıran basit ama etkili bir algoritmadır.

Makine Öğrenmesi (Machine Learning): Bilgisayarların açıkça programlanmadan, verilerden öğrenerek tahmin veya karar verebilmesini sağlayan yapay zeka dalıdır.

Naive Bayes: Bayes teoremini kullanarak olasılık temelli sınıflandırma yapan, özellikler arasında bağımsızlık varsayımı yapan algoritmasıdır.

Nem (Humidity, Hum): Ortamda veya toprakta bulunan su buharı miktarıdır.

Oylama Sınıflandırıcısı (Voting Classifier): Birden fazla farklı modelin tahminlerini birleştirerek nihai sonucu belirleyen topluluk yöntemidir.

PCA (Principal Component Analysis): Yüksek boyutlu verilerde, bilgiyi koruyarak boyutu küçültmek için kullanılan istatistiksel yöntemidir.

pH (Potential of Hydrogen): Toprağın asitlik veya alkalilik derecesini gösteren ölçüdür.

Rastgele Orman (Random Forest, RF): Çok sayıda karar ağacının sonuçlarını birleştirerek sınıflandırma veya regresyon yapan topluluk algoritmasıdır.

Yağış (Rainfall): Belirli bir süre zarfında düşen yağmur miktarıdır.

Yapay Sinir Ağları (Artificial Neural Networks, ANN): İnsan beynindeki nöron yapısını taklit eden, çok katmanlı öğrenme modelleridir.

GİRİŞ

Tarım sektörü, gıda üretiminin ötesinde sosyal refah, çevresel sürdürülebilirlik ve ekonomik kalkınmada kritik rol oynar. Hindistan gibi geniş tarım arazilerine sahip ülkelerde doğru ürün seçimi, hem verimlilik hem de gıda güvenliği açısından hayati önemdedir. Ancak yağış, sıcaklık, toprak özellikleri gibi değişkenler öngörülemez olduğundan geleneksel yöntemler yetersiz kalır. Bilgiye erişim kısıtlılığı, yanlış seçimlerle ekonomik kayıplara yol açabilir. Yapay zekâ ve makine öğrenmesi (ML) bu belirsizlikleri azaltarak karar kalitesini yükseltir.

Çalışmanın amacı, Hindistan'daki çiftçilerin doğru mahsul seçimi yaparken karşılaştıkları belirsizlikleri ve zorlukları azaltmaktır. Bu amaçla, elde edilen veritabanından alınan ve mahsul bilgileri, toprak bileşimi, sıcaklık, pH, yağış gibi verileri içeren veri seti üzerinde Naive Bayes, Gradient Boosting, Yapay Sinir Ağları, Random Forest ve Oylama Sınıflandırıcısı gibi algoritmalar kullanılmıştır. Hedef, tarımsal üretkenliği artıracak, yanlış mahsul seçimlerini azaltacak ve sürdürülebilir tarım politikalarına katkı sağlayacak bir karar destek sistemi geliştirmektir.

Genel problem, çiftçilerin geleneksel yöntemlere dayalı kararlarının çevresel değişkenler karşısında yetersiz kalması, bunun da düşük verim ve gelir kaybına yol açmasıdır. Yanlış ürün seçimi; su israfı, gereksiz gübreleme, toprak yorgunluğu ve hastalık yayılımı gibi sorunlar doğurur. Çözüm, geniş ve çok boyutlu tarımsal verilerin işlenmesi, analiz edilmesi ve sınıflandırma algoritmalarıyla en uygun mahsulün belirlenmesidir.

Tezin sınırlılıkları, yalnızca belirli bir veri setine ve Hindistan'daki tarımsal koşullara dayanmasıdır. Diğer ülkelerde veya farklı veri kaynaklarında doğrulama yapılmamıştır. Ayrıca, model performansı kullanılan veri kalitesi, eksik değer yönetimi ve ön işleme adımlarına bağlıdır. Algoritmalar seçilirken veri setinin yapısına uygun olanlar tercih edilmiştir, fakat bağımlı özelliklerin fazla olduğu durumlarda bazı yöntemlerin performansı düşebilir.

Veriye dayalı tarım, üretim artışı ve kaynak verimliliği sağlar; besin maddelerinin doğru analizi gereksiz gübreleme ve sulamayı önler. Algoritmaların başarısı, teknik doğruluğun yanında kullanıcı dostu arayüzler, yerel uyum ve dijital

okuryazarlık ile de ilişkilidir. Naive Bayes hızlı ve yüksek doğrulukta sonuç verirken, ensemble yöntemleri genelleme kabiliyetini artırır.

Sonuç olarak, bu çalışma tarımsal karar süreçlerinde veri ve teknoloji tabanlı dönüşümün çerçevesini sunar. Daha geniş coğrafyalarda ve farklı veri setlerinde yapılacak testler, modelin genelleme yeteneğini güçlendirecek ve sürdürülebilir kalkınma hedeflerine katkı sağlayacaktır.

1. MAKİNE ÖĞRENİMİ ve TEZ HAKKINDA

Günümüz teknolojisinin en dikkat çekici ve hızla gelişen alanlarından biri olan Makine Öğrenimi (Machine Learning), bilgisayar sistemlerinin dışarıdan açıkça programlanmadan, veriler aracılığıyla öğrenmesini ve bu öğrenim sonucunda çeşitli görevleri yerine getirebilmesini sağlayan bir yapay zeka dalıdır. Geleneksel programlamanın aksine, makine öğrenimi sistemleri belirli kuralları önceden yazılmış algoritmalarla uygulamak yerine, büyük veri kümeleri üzerinde analizler yaparak örüntüleri (pattern) keşfeder ve bu örüntüler doğrultusunda tahminlerde bulunabilir ya da karar verebilir.

Makine öğrenimi, istatistik, matematik, bilgisayar bilimi ve bilgi teorisi gibi çeşitli disiplinlerin kesişim noktasında yer alır. Özellikle son yıllarda donanım kapasitesinin artması, büyük veri analitiği uygulamalarının yaygınlaşması ve algoritmaların daha da gelişmesiyle birlikte, bu alan sağlık, finans, tarım, otomotiv, güvenlik, eğitim ve pazarlama gibi birçok sektörde devrim niteliğinde değişikliklere yol açmıştır.

1.1. Makine Öğrenmesine Giriş

Tarım sektörü, dünya ekonomisinin en temel ve hayati parçalarından birini oluşturmaktadır. Gıda üretiminin artırılması ve kaynakların verimli kullanılabilmesi için teknolojinin tarıma entegrasyonu büyük önem taşımaktadır. Geleneksel tarım yöntemleri, doğal kaynakları koruyarak daha yüksek verim elde etmeyi hedeflerken, bu yöntemler genellikle sınırlı veri analizine dayanmaktadır. Makine öğrenmesi (ML) gibi ileri düzey algoritmalar, tarımda hem verimliliği artırmak hem de çevresel sürdürülebilirliği sağlamak adına büyük bir potansiyel taşımaktadır. Bu teknolojiler, geniş veri kümelerinin analiz edilmesine olanak vererek, doğru ürün seçimlerini yapmak için çiftçilere güçlü bir araç sunmaktadır.

Makine öğrenmesinin tarımda kullanımının en yaygın alanlarından biri, ürün seçiminin optimize edilmesidir. Çiftçilerin hangi ürünleri hangi koşullarda yetiştireceklerini seçerken karşılaştıkları zorluklar, çevresel değişkenlerin yüksekliği ve sürekli değişen iklim koşullarıdır. Makine öğrenmesi, bu değişkenlerin etkilerini analiz ederek daha bilinçli ve doğru kararlar alınmasını sağlamaktadır. Makine öğrenmesi,

verilerden öğrenme sürecine dayanır ve belirli görevleri yerine getirmek için programlama gerektirmeyen bir yapay zeka teknolojisidir. Tarımda kullanılan makine öğrenmesi algoritmalarını daha iyi anlayabilmek için, bunların üç ana kategorisini ve her birinin tarımda nasıl uygulanabileceğini incelemek faydalı olacaktır.

Gözetimli öğrenme, en yaygın kullanılan makine öğrenmesi türlerinden biridir. Bu algoritmalarda, bir model etiketlenmiş verilerle eğitilir. Bu, giriş verileri (örneğin, toprak pH değeri, nem oranı, sıcaklık vb.) ve bunlara karşılık gelen çıktılar (örneğin, ürün verimi) arasındaki ilişkiyi öğrenmeye dayanır. Gözetimli öğrenme, özellikle geçmişte toplanan verilerle gelecekteki olayları tahmin etmek için kullanılır. Tarımda doğrusal regresyon veya karar ağaçları gibi algoritmalar kullanılarak, geçmiş yıllarda toplanan çevresel verilerle hangi ürünlerin hangi koşullarda daha yüksek verim verdiği analiz edilebilir. Örneğin, belirli bir toprak tipine ve iklim koşullarına sahip bir bölgede, bu veriler doğrultusunda hangi tarım ürünlerinin daha iyi yetiştiği tahmin edilebilir. Destek vektör makineleri (SVM) gibi daha karmaşık gözetimli algoritmalar da, büyük veri kümelerinde daha karmaşık ilişkileri öğrenebilir ve belirli toprak türleri ve iklim koşulları altında en uygun ürünleri tahmin etmek için kullanılabilir.

Gözetimsiz öğrenme, etiketlenmemiş verilerle çalışarak veri kümesindeki gizli yapıları keşfeder. Tarımda bu algoritmalar, verilerin gruplandırılması ve belirli bölgelerde hangi ürünlerin daha verimli yetiştiğinin belirlenmesi amacıyla kullanılabilir. Kümeleme algoritmaları (örneğin, k-means) ile belirli toprak türleri ve çevresel koşullara göre farklı bölgelerde benzer özelliklere sahip alanlar gruplandırılabilir. Bu kümeler, o bölgeye özgü ürünlerin belirlenmesinde kullanılabilir. Ayrıca, anlamlı bağlantılar veya yapısal desenler bulmaya yardımcı olan PCA (Principal Component Analysis) gibi tekniklerle, toprak ve iklim verilerindeki gizli yapılar belirlenebilir. Bu analizler, hangi ürünlerin belirli çevresel koşullarda daha verimli büyüdüğüne dair bilgiler sağlar.

Pekiştirmeli öğrenme, bir ajanın çevresiyle etkileşimde bulunarak ödüller alıp eylem stratejileri geliştirdiği bir öğrenme türüdür. Tarımda, bu yöntem, sürekli değişen çevresel faktörlere dayanarak strateji geliştiren sistemler için kullanılır. Örneğin, sulama sistemlerinin optimizasyonu, toprak nemi, sıcaklık ve yağış gibi verilerle beslenen bir makine öğrenmesi modeli tarafından gerçekleştirilebilir. Bu model, her gün için su

miktarını optimize ederek sulama sistemlerini kontrol edebilir ve böylece su israfını önleyebilir. Ayrıca, pekiştirmeli öğrenme, toprak analizine dayalı olarak hangi gübrelerin ve ne kadar kullanılacağını öğrenebilir. Bu süreç, verimliliği artırırken çevreye olan kimyasal zararları da minimize eder.

Makine öğrenmesinin başarısı, doğru ve kapsamlı veriye dayanır. Tarımda kullanılan veriler, çeşitli kaynaklardan toplanabilir ve bu verilerin kalitesi, algoritmaların doğruluğu üzerinde doğrudan etki eder. Bu veriler, çevresel veriler (sıcaklık, nem oranı, rüzgar hızı, yağış miktarı ve güneş ışınımı gibi iklimsel faktörler), toprak verileri (pH seviyesi, toprak nemi, organik madde oranı, mineral içeriği, toprak sıcaklığı ve yapısı), biyolojik veriler (ürün türleri, bitki sağlığı, verim oranları, hastalıklar, zararlılar ve uygulanan pest kontrolü) ve uzaktan algılama verileri (uydu ve drone görüntüleri ile tarım alanlarının sağlık durumu ve gelişim seviyeleri) gibi çeşitli kategorilerde toplanabilir. Bu veriler, veri madenciliği teknikleri kullanılarak işlenebilir ve daha sonra makine öğrenmesi algoritmalarına beslenebilir. Örneğin, uydu verileri ve sensörlerden alınan bilgilerle, tarım alanlarındaki bitkilerin büyüme durumu analiz edilebilir. Bu veriler, doğru ürün seçiminde yardımcı olabilir.

1.2. Makine Öğrenmesi

Makine öğrenmesi, bilgisayarların açıkça programlanmadan, verilerden öğrenerek karar verme ve tahmin yapma yeteneklerini geliştirmeyi amaçlayan bir yapay zekâ alt alanıdır. Modern bilgisayar bilimlerinin en aktif ve üretken disiplinlerinden biri hâline gelen makine öğrenmesi, günümüzde tarımdan sağlığa, finanstan siber güvenliğe kadar pek çok sektörde devrim niteliğinde uygulamalara zemin hazırlamıştır. Ancak, makine öğrenmesinin kökenleri bugünkü kadar sofistike olmasa da, oldukça eski dönemlere dayanmaktadır. Makine öğrenmesinin fikirsel temelleri 1950'li yıllarda Alan Turing'in "Bilgisayarlar düşünebilir mi?" sorusunu sormasıyla atılmıştır. Turing, makinelerin zekice davranış sergileyebilme potansiyelini sorgulamış ve bu yaklaşım, yapay zekâ araştırmalarına ilham vermiştir.

Makine öğrenmesi teriminin literatürde ilk kez kullanılması ise Arthur Samuel'in 1959 yılında IBM'deki çalışmaları sırasında gerçekleşmiştir. Samuel, makine öğrenimini "bilgisayarların açıkça programlanmadan öğrenme yeteneği"

olarak tanımlamıştır. Geliştirdiği dama oynayan yazılım, kendi kendine oynayarak stratejiler geliştirmiş ve bu, erken dönemlerde makine öğreniminin uygulamalı bir örneği olarak kabul edilmiştir. Bu dönemde geliştirilen sistemler sınırlı işlem gücü ve düşük veri kapasitesi nedeniyle temel düzeyde kalmış olsa da, bu araştırmalar makine öğreniminin doğuşunu simgelemektedir.

1980'ler ve 1990'lar, makine öğrenmesi tarihinde önemli bir dönüşüm dönemidir. Bu süreçte, istatistiksel öğrenme kuramları ve yapay sinir ağlarına yönelik yoğun akademik ilgi doğmuştur. Özellikle 1986 yılında Rumelhart, Hinton ve Williams tarafından geliştirilen geri yayılım (backpropagation) algoritması, çok katmanlı yapay sinir ağlarının eğitilmesini mümkün kılmış ve sinir ağları araştırmalarına yeni bir boyut kazandırmıştır. Bu gelişme, makine öğrenmesinin yalnızca teorik değil, aynı zamanda uygulamalı başarılar elde etmesini sağlamış ve görüntü işleme, ses tanıma gibi alanlarda kullanılmaya başlamıştır. Bu yıllarda geliştirilen SVM (Destek Vektör Makineleri), K-En Yakın Komşu (K-NN) ve Karar Ağaçları gibi algoritmalar da bugünün makine öğrenmesi altyapısını oluşturan temel taşlar arasında yer almıştır.

2000’li yıllarla birlikte veri üretiminin büyük bir ivmeyle artması, makine öğrenmesinde bir paradigma değişikliğine yol açmıştır. Özellikle internetin yaygınlaşması, sosyal medyanın gelişimi ve dijitalleşen endüstriler sayesinde büyük veri (big data) kavramı ortaya çıkmış; bununla paralel olarak da yüksek işlem gücüne sahip donanımların kullanımı yaygınlaşmıştır. Bu gelişmeler, daha büyük ve karmaşık veri setlerinin işlenebilmesini mümkün kılmış ve klasik makine öğrenmesi yöntemlerinin ötesine geçilmesini sağlamıştır. Artık veri analizi ve öğrenme süreçleri, daha karmaşık yapılarla yürütülmekte, verideki örüntüler çok katmanlı yapılarda keşfedilebilmektedir.

Bu dönemle birlikte derin öğrenme (deep learning) olarak adlandırılan yapay sinir ağı mimarileri ön plana çıkmıştır. Derin öğrenme, çok sayıda katmandan oluşan sinir ağlarının, veriden soyut düzeyde özellikler çıkararak öğrenmesini mümkün kılan bir yaklaşımdır. Özellikle görüntü tanıma (image recognition), nesne tespiti, doğal dil işleme (NLP), otomatik konuşma tanıma ve özerk sürüş gibi alanlarda çığır açan sonuçlar elde edilmiştir. Convolutional Neural Networks (CNN) ve Recurrent Neural

Networks (RNN) gibi mimariler sayesinde, makine öğrenmesi uygulamaları görsel, metinsel ve zamansal veriler üzerinde olağanüstü başarılar sergilemiştir.

2010'lu yıllar ve sonrasındaki süreç, makine öğrenmesinin günlük yaşamla daha fazla bütünleştiği bir dönem olmuştur. Otomatik öneri sistemleri (örneğin Netflix ve YouTube'da önerilen içerikler), sanal asistanlar (Alexa, Siri), çevrimiçi çeviri hizmetleri (Google Translate) ve kişiselleştirilmiş reklamcılık gibi hizmetlerde makine öğrenmesi rutin bir altyapı olarak kullanılmaya başlanmıştır. Ayrıca tıp alanında görüntüleme verilerinden teşhis üretme, finans sektöründe kredi risk analizi, siber güvenlik alanında anomali tespiti ve daha birçok sektörde makine öğrenmesi, stratejik karar destek sistemlerinin temelini oluşturmaktadır. Bu gelişmelerin ardında, model performansını artıran güçlü algoritmalar kadar, veri kalitesinin yükselmesi ve açık kaynaklı yazılım kütüphanelerinin (TensorFlow, PyTorch, Scikit-learn gibi) yaygınlaşması da büyük rol oynamıştır.

Makine öğrenmesinin bu tarihsel evriminde önemli teorik katkılar da söz konusudur. İstatistiksel öğrenme kuramları, model karmaşıklığı ve genelleme yeteneği arasındaki ilişkiyi analiz eden çerçeveler sunmuştur. Özellikle Vapnik-Chervonenkis (VC) boyutu, yapay zekâ modellerinin sınıflandırma başarısını açıklamak için geliştirilmiş önemli bir ölçüttür. Ayrıca Bayesci yaklaşımlar, olasılık temelli modelleme yapısıyla belirsizlik altında karar vermeyi olanaklı kılmış ve sağlık, meteoroloji, ekonomi gibi yüksek riskli alanlarda sıklıkla tercih edilen yöntemler hâline gelmiştir. Bu teorik altyapı, makine öğrenmesi uygulamalarının sadece veri madenciliği ile sınırlı kalmamasını, aynı zamanda karar teorisi, bilgi kuramı ve istatistik gibi disiplinlerle etkileşimli gelişmesini sağlamıştır.

Bugün makine öğrenmesi, yalnızca teknolojik bir araç değil; veriyle düşünebilen, örüntüler üzerinden karar alabilen, deneyimden öğrenebilen akıllı sistemlerin temelini oluşturan stratejik bir paradigma hâline gelmiştir. Bu bağlamda, çalışmamızda kullanılan makine öğrenmesi algoritmaları da bu tarihsel ve kavramsal gelişmelerin ürünüdür. Bu tez kapsamında tercih edilen algoritmalar, veri setinin özelliklerine ve problem tanımına göre seçilmiş ve her biri belirli avantajlar sunacak şekilde uygulanmıştır. Naive Bayes gibi istatistiksel yaklaşımlar, doğruluğu yüksek olan fakat açıklanabilirliği sınırlı modellerle desteklenmiş; karar ağaçları gibi

yorumlanabilir modeller, modelin açıklanabilirliğini artırmak amacıyla devreye alınmıştır. Ayrıca bazı senaryolarda ensemble yöntemleri (bagging, boosting gibi) kullanılarak daha güçlü tahmin sonuçları elde edilmiştir.

Makine öğrenmesinin bu evrimi, yalnızca geçmişten bugüne bir başarı hikâyesi değil, aynı zamanda geleceğe dönük büyük bir potansiyeli de temsil etmektedir. Gelişen donanım altyapısı, artan veri miktarı ve gelişmiş algoritmalar sayesinde, önümüzdeki yıllarda çok daha karmaşık problemler, çok daha hassas şekilde çözülebilecektir. Tarım gibi geleneksel alanların dijitalleşmesi, veri odaklı karar destek sistemlerinin yaygınlaşması ve sürdürülebilir kalkınma hedeflerine katkı sağlamak, makine öğrenmesinin hem akademik hem de endüstriyel bağlamda daha da önem kazanmasına neden olacaktır.

1.3. Makine Öğrenmesi ve Tarım

Tarımda makine öğrenmesi uygulamaları, yukarıda belirtilen dört temel öğrenme türünün çeşitli varyasyonlarıyla gerçekleştirilebilir. Örneğin, denetimli öğrenme doğru ürün seçiminde, denetimsiz öğrenme toprak analizi veya bölgesel ürün uyumluluğu değerlendirmelerinde, yarı denetimli öğrenme etiketsiz tarım arazisi görüntülerinin işlenmesinde, takviyeli öğrenme ise dinamik sistemlerde, örneğin iklim adaptasyonu gibi alanlarda kullanılabilir. Bu sayede çiftçiler, hangi ürünü ne zaman ve nerede ekeceklerine daha isabetli kararlar verebilir, su ve gübre kullanımını optimize edebilir ve hastalık ile zararlı tespiti konusunda erken uyarı sistemlerine sahip olabilirler.

21. yüzyılda dünya nüfusunun hızla artması, gıda talebinin de buna paralel olarak büyümesine yol açmaktadır. Bu durum, sınırlı doğal kaynaklar üzerinde artan baskı oluşturmakta ve özellikle tarım sektörünü daha verimli, sürdürülebilir ve akıllı çözümler üretmeye zorlamaktadır. Geleneksel tarım yöntemleri, değişen iklim koşulları, arazi bozulmaları ve kaynak kıtlığı gibi küresel sorunlara tek başına yanıt veremez hale gelmiştir. Bu noktada, makine öğrenmesi (ML) gibi yapay zeka destekli teknolojiler, tarımda devrim yaratacak potansiyele sahip bir araç olarak öne çıkmaktadır. Veriye dayalı karar destek sistemleri oluşturarak, doğru ürün seçimi,

zamanında sulama, erken hastalık tespiti ve optimum gübreleme gibi uygulamalarla daha az kaynakla daha fazla üretim yapılmasını mümkün kılmaktadır.

Makine öğrenmesi, bilgisayar sistemlerinin geçmiş verilere dayanarak örüntüler tanıması, tahminlerde bulunması ve zamanla kendi performansını iyileştirmesi prensibine dayanır. Tarım gibi çok değişkenli, doğayla iç içe ve öngörülemez dış etkenlere maruz kalan bir alanda bu teknoloji, belirsizlikleri azaltmak ve üreticilere bilimsel tabanlı öngörüler sağlamak açısından kritik bir rol oynamaktadır. ML sistemleri, tarımsal verileri analiz ederek neyi (ürün seçimi), nerede (lokasyon bazlı analiz) ve ne zaman (optimum ekim/hasat zamanlaması) gerçekleştirmek gerektiğini belirleyebilir. Bu üç soru, tarımsal karar alma süreçlerinin temelini oluşturur ve makine öğrenmesi bu sorulara matematiksel, model tabanlı ve öğrenen sistemlerle yanıt verir.

Makine öğrenmesi teknikleri, tarımda farklı görevlerde çeşitli amaçlara hizmet edecek şekilde uyarlanabilir. Tarımda yaygın olarak kullanılan bazı ML algoritmaları arasında Random Forest, SVM (Support Vector Machine), K-Means Clustering, LSTM (Long Short-Term Memory) ve CNN (Convolutional Neural Networks) bulunmaktadır. Random Forest, çok sayıda karar ağacı üzerinden tahmin yapan bir sınıflandırma algoritmasıdır ve ürün verim tahmini ile zararlı analizi için kullanılır. SVM, verileri düzlemler aracılığıyla sınıflandırır ve toprak-sıcaklık ilişkisine göre ürün sınıflaması yapar. K-Means Clustering, verileri benzerliklerine göre gruplar ve toprak tipi veya bölge analizi için kullanılır. LSTM, zaman serisi verileriyle çalışan bir yapay sinir ağı türüdür ve iklim ile hava durumu öngörüsü için kullanılır. CNN ise görüntü verilerini işleyerek analiz yapar ve bitki yaprak hastalıklarının tanısı için kullanılır.

Makine öğrenmesi, performansını büyük ölçüde veri kalitesine ve çeşitliliğine borçludur. Tarımda kullanılan veri türleri arasında iklim verileri (sıcaklık, nem, yağış, rüzgar verileri), toprak verileri (pH, organik madde, azot seviyesi, tuzluluk, nem), görüntü verileri (drone, uydu ve cep telefonu ile alınan görüntüler), zaman serisi verileri (mevsimsel verim kayıtları, geçmiş ürün tercihleri, fiyat dalgalanmaları) ve saha verileri (gübrelere verilen tepki, sulama geçmişi, ekim/hasat zamanı bilgileri)

bulunmaktadır. Bu veriler, farklı algoritmalara göre temizlenir, etiketlenir, ölçeklenir ve modele hazır hale getirilir.

Tarımda ürün seçimi, yalnızca çiftçinin bireysel tercihi değil, aynı zamanda iklim, toprak, pazar koşulları ve su kaynakları gibi birçok parametreye bağlıdır. ML algoritmaları, çok sayıda değişkeni aynı anda analiz ederek en yüksek verim sağlayacak ürün türünü belirlemede kullanılır. Örneğin, Kuzeydoğu Anadolu bölgesinde yapılan analizde, toprak tuzluluğu yüksek olan alanlarda makine öğrenmesi destekli model, arpa ve çavdar gibi tuz toleranslı türlerin önerilmesini sağlamıştır. Ege bölgesinde ise sıcaklık, nem ve pH verilerine dayanarak pamuk ve zeytin gibi ürünlerin bölgeye daha uygun olduğu tespit edilmiştir. Bu tür modeller, bölgesel bazda özelleştirilmiş tarım politikaları oluşturulmasına da katkı sunmaktadır.

Makine öğrenmesi uygulamaları sayesinde gübre ve ilaç kullanımında %20-30'a varan tasarruf sağlanabilir. Ürün hastalıklarının erken teşhisi ile %40'a varan verim kaybı önlenebilir. Zamanında ve yeterli sulama ile hem enerji hem de su tüketimi azaltılabilir. Hasat planlaması ile iş gücü optimizasyonu sağlanır. Bu kazanımlar yalnızca büyük ölçekli tarım işletmeleri için değil, küçük aile çiftçileri için de uygulanabilir hale geldikçe, gıda güvenliği ve kırsal kalkınma için ciddi faydalar doğacaktır.

Gerçek dünya uygulamaları arasında Microsoft AI for Earth, Kenya'da yapay zeka ile toprak sağlığı analizleri yaparak ürün önerisi sunmaktadır. IBM Watson Decision Platform for Agriculture, Amerika'da binlerce dönüm arazide sensör verileriyle bitki gelişimini izlemektedir. Plantix uygulaması, Hindistan'da çiftçilerin cep telefonları ile yaprak resmi çekip hastalık tanısı almasına olanak tanımaktadır. TARLA, TARBİL ve TÜBİTAK destekli projelerde Türkiye'ye özgü toprak ve iklim verilerine dayalı algoritmalar geliştirmektedir.

Makine öğrenmesi teknolojilerinin tarımda yaygınlaşmasının önünde bazı engeller ve dikkat edilmesi gereken etik sorunlar bulunmaktadır. Veri mahremiyeti, çiftçilerin veri haklarının korunması ve verilerin izinsiz paylaşılmaması gerekmektedir. Teknolojiye erişim uçurumu, gelişmiş bölgeler ile kırsal alanlar arasında ciddi bir dijital uçurum yaratmaktadır. Tarafsızlık ve adalet, bu teknolojilerin uygulanmasında göz önünde bulundurulması gereken önemli unsurlardır.

1.4. Tez Hakkında

Bu tez, tarım sektöründe doğru mahsul seçimi sürecini veri temelli hale getirmek amacıyla makine öğrenmesi algoritmalarını kullanmaktadır. Kullanılan veri seti; toprak besin değerleri (Azot, Fosfor, Potasyum), iklimsel değişkenler (sıcaklık, nem, yağış) ve toprak pH değeri gibi parametreleri içermektedir. Bu veriler, farklı makine öğrenmesi algoritmaları (Naive Bayes, Gradient Boosting, Yapay Sinir Ağları, Random Forest ve Oylama sınıflandırıcı) ile analiz edilerek belirli bir bölge ve koşullar için en uygun mahsul önerilmektedir. Amaç, çiftçilerin belirsizlik altında verdikleri kararları bilimsel öngörülerle destekleyerek verimliliği ve sürdürülebilirliği artırmaktır.

1.5. Tezin Problemi

Geleneksel yöntemlerle mahsul seçimi, iklim değişkenliği, toprak yapısındaki farklılıklar ve bilgiye erişim eksikliği gibi faktörler nedeniyle çoğu zaman yanlış kararlar doğurmaktadır. Bu durum; verim kaybı, gelir düşüşü, su ve gübre israfı gibi hem ekonomik hem de çevresel olumsuzluklara yol açmaktadır. Ayrıca, karmaşık ve çok değişkenli tarımsal verinin manuel yöntemlerle analiz edilmesi mümkün olmadığından, üreticiler çoğunlukla sezgisel ve deneyime dayalı seçimler yapmakta; bu da tarımsal verimlilikte istikrarsızlığa sebep olmaktadır.

1.6. Tezin Amacı

Tezinin temel amacı, veri analitiği ve Makine Öğrenmesi teknikleri ile belirli bir arazi için en uygun mahsulü doğru ve hızlı şekilde belirleyebilecek bir model geliştirmektir. Bu model; çiftçilere doğruluk oranı yüksek, bilimsel tabanlı öneriler sunarak yanlış ürün seçimini azaltmayı, kaynak kullanımını optimize etmeyi ve ekonomik getiriyi artırmayı hedeflemektedir. Ayrıca, gıda güvenliğine katkı sağlamak ve sürdürülebilir tarım uygulamalarını yaygınlaştırmak da tezin amaçları arasındadır.

1.7. Tezin Kapsamı

Tez; veri toplama, ön işleme, modelleme, değerlendirme ve öneri üretme aşamalarını kapsamaktadır:

Veri Toplama: N, P, K, sıcaklık, nem, pH, yağış ve mahsul bilgilerini içeren tarımsal veri setinin temini.

Ön İşleme: Eksik/veri hatalarının giderilmesi, normalizasyon ve kodlama işlemleri.

Modelleme: Naive Bayes, KNN, Random Forest, Gradient Boosting, Yapay Sinir Ağları ve Oylama sınıflandırıcı gibi algoritmalarla model eğitimi.

Değerlendirme: Eğitim/test ayrımı yapılarak doğruluk, kesinlik, duyarlılık ve F1 skoru gibi metriklerle performans analizi.

Karar Desteği: En iyi performans gösteren modelin çiftçilere mahsul önerisi sunması.

1.8 Tezin Önemi

Bu tez, tarım sektöründe **verimlilik artışı, kaynak optimizasyonu ve gıda güvenliği** açısından stratejik bir değere sahiptir.

- **Ekonomik Katkı:** Doğru mahsul seçimi ile çiftçi gelirlerinde artış.
- **Çevresel Fayda:** Su, gübre ve diğer girdilerin gereksiz kullanımının önlenmesiyle çevresel etkilerin azaltılması.
- **Sosyal Etki:** Kırsal bölgelerde yaşam standardının yükselmesi ve bilgiye dayalı tarım kültürünün yaygınlaşması.
  Bu yönleriyle tez, hem bireysel hem de toplumsal düzeyde sürdürülebilir kalkınma hedeflerine hizmet etmektedir.

1.9. Tezin Sınırlılıkları

- **Veri Kaynağı Kısıtı:** Kullanılan veri seti yalnızca Hindistan'a özgüdür; farklı bölgelerde geçerliliği test edilmemiştir.

- **Veri Kalitesine Bağımlılık:** Modelin performansı, veri ön işleme sürecindeki kaliteye ve veri bütünlüğüne doğrudan bağlıdır.
- **Algoritma Varsayımları:** Bazı modeller (ör. Naive Bayes) özelliklerin bağımsız olduğu varsayımına dayanır; bu durum gerçek dünyada her zaman geçerli olmayabilir.
- **Genelleme Sorunu:** Sonuçların farklı iklim, toprak ve sosyoekonomik koşullara uyarlanması için ek çalışmalar gereklidir.

1.10. K-En Yakın Komşu

K-NN Yakın Komşular (K-Nearest Neighbor) algoritması, denetimli bir makine öğrenmesi algoritmasıdır ve sınıflandırma problemlerinde yaygın olarak kullanılmaktadır. Bu algoritma, bir veri noktasının sınıfını belirlemek için, komşu sayısı parametresi belirtilerek, veri setindeki noktaların her birinin kendisine en yakın olan komşu noktalarını baz alır. K-En Yakın Komşular algoritması, belirli bir veri noktasının sınıfını tahmin etmek için, o noktaya en yakın K komşusunun sınıflarını kullanır. Bu yaklaşım, veri noktalarının benzerliklerine dayalı olarak sınıflandırma yapma yeteneği sunar. Algoritma, veri setindeki eğitim verisi ile eğitilir ve daha sonra eğitilen bu veri seti üzerinden yaptığı çıkarımları test verisi üzerinde sınıflandırma veya tahmin yapmak için kullanır. Eğitim aşamasında, algoritma, her bir veri noktasının özelliklerini ve bunlara karşılık gelen sınıflarını öğrenir. Test aşamasında ise, yeni bir veri noktası verildiğinde, algoritma bu noktanın en yakın K komşusunu belirler ve bu komşuların sınıflarını dikkate alarak tahmin yapar. Veriler arasındaki mesafeyi ölçmek için çeşitli mesafe ölçütleri kullanılmaktadır. En yaygın olarak kullanılan mesafe ölçütlerinden biri Euclidean mesafesidir. Bunun yanı sıra, Manhattan mesafesi gibi diğer ölçütler de kullanılabilir. Euclidean mesafesi, iki veri noktası arasındaki düz bir çizgi boyunca ölçülen mesafeyi ifade ederken, Manhattan mesafesi, iki nokta arasındaki mesafeyi, yalnızca yatay ve dikey hareketlerle hesaplar. Aşağıda, Manhattan mesafesinin matematiksel ifadesi verilmiştir:

$$D = \Sigma \, |x\,i - y\,i| \qquad \textit{(Denklem 1.)}$$

Bu denklemde, (D) iki veri noktası arasındaki Manhattan mesafesini, (x_i) ve (y_i) ise karşılaştırılan veri noktalarının özelliklerini temsil etmektedir (Tan vd., 2016). K-En Yakın Komşular algoritması, basit yapısı ve etkili sonuçları sayesinde, birçok uygulama alanında tercih edilmektedir. Özellikle, sınıflandırma problemlerinde, veri setinin boyutu ve karmaşıklığı arttıkça, bu algoritmanın performansı da dikkat çekici bir şekilde artmaktadır.

1.11. Naive Bayes

Naive Bayes, olasılıksal bir sınıflandırıcıdır ve makine öğrenmesi alanında yaygın olarak kullanılan bir yöntemdir. Bu sınıflandırmaları yaparken, Bayes teoreminin koşullu olasılık prensibini kullanır. Naive Bayes algoritması, her bir özelliğin sınıf etiketi üzerindeki etkisini bağımsız olarak değerlendirir. Bu yaklaşım, her bir özelliğin mevcut olmasının diğer özellikleri etkilemediği varsayımına dayanır. Bu nedenle, "naive" (saf) terimi, algoritmanın bu basitleştirici varsayımından gelmektedir. Sınıflandırma sırasında, Naive Bayes, veri setindeki özelliklerin değerlerini gözlemleyerek her sınıfın olasılığını hesaplar. Özellikle, her bir sınıf için, o sınıfa ait olma olasılığı, o sınıfın özelliklerinin gözlemlenen değerleri ile birlikte değerlendirilir. Ardından, en yüksek olasılığa sahip sınıfı tahmin eder. Bu süreç, genellikle hızlı ve verimli bir şekilde gerçekleştirilir, bu da Naive Bayes algoritmasını kolay uygulanabilir ve genellikle güzel sonuçlar veren bir yöntem haline getirir. Özellikle metin sınıflandırma, spam tespiti ve duygu analizi gibi alanlarda sıkça tercih edilmektedir.

Naive Bayes algoritmasının temel mantığı, belirli bir örnek kümesi (X = (x_1, x_2, \ldots, x_n)) ve sınıf kümesi (C = (C_1, C_2, \ldots, C_m)) üzerinden çalışır. Amaç, belirli bir sınıf için (P(C_i | X)) olasılığını maksimize etmektir. Bu bağlamda, (P(C_i | X)) olasılığının maksimize edilmesine maksimum sonsal hipotez (maximum posteriori hypothesis) denir.

Bayes teoremi, aşağıdaki gibi ifade edilir:

$$P\,(C\,i\ \ |X\,) = P\,(C\,i\,)P\,(X\,|C\,i\,)\,P\,(X\,)\quad \text{(Denklem 2.)}$$

Bu denklemde, (P(C_i)) sınıfın önsel olasılığını, (P(X | C_i)) ise verilen sınıf için gözlemlenen özelliklerin olasılığını temsil eder. (P(X)) ise gözlemlenen özelliklerin genel olasılığıdır. Naive Bayes algoritması, bu olasılıkları hesaplayarak, her bir sınıf için en yüksek olasılığa sahip olanı seçer.

Ancak, Naive Bayes algoritmasının bazı sınırlamaları da bulunmaktadır. Özellikler arasındaki gerçek bağımlılıkları göz ardı ettiği için, bazı durumlarda performansı düşük olabilir. Özellikle, özelliklerin birbirine bağımlı olduğu durumlarda, bu bağımsızlık varsayımı, sınıflandırma sonuçlarını olumsuz etkileyebilir. Bu nedenle, Naive Bayes algoritması, bağımsızlık varsayımının geçerli olduğu durumlarda en iyi performansı gösterirken, bağımlı özelliklerin bulunduğu durumlarda dikkatli bir şekilde kullanılmalıdır.

1.12. Yapay Sinir Ağları

Yapay Sinir Ağları (ANN), perceptron adı verilen hücrelerden oluşan ve karmaşık problemleri çözmek için insan beynindeki nöronların işleyişini taklit eden bir simülasyon modelidir. Bu perceptronlar, diğer bazı perceptronlardan girdi alan, bu girdiler üzerinde hesaplamalar yapan ve diğer girdileri besleyen hesaplama birimleridir (McCulloch, 1990). Yapay sinir ağları, girdi değişkenleri ve çıktı değişkenleri arasındaki matematiksel ilişkileri "öğrenme" yeteneğine sahip bir algoritmadır. Bu öğrenme süreci, ağın belirli bir görev veya problem üzerinde performansını artırmak amacıyla gerçekleştirilir.

Yapay sinir ağlarında, istenirse girdi ve çıktı katmanları arasında bir veya daha fazla gizli katman eklenebilir. Gizli katmanlar, ağın karmaşıklığını artırarak daha derin ve anlamlı ilişkilerin öğrenilmesine olanak tanır. Bu katmanlar arasında çeşitli düğümler bulunur. Her düğüm, giriş özelliklerinden aldığı sinyalleri alır, bu sinyalleri ağırlıklarla çarpar ve bir aktivasyon fonksiyonuna (genellikle sigmoid veya ReLU gibi) gönderir. Aktivasyon fonksiyonu, düğümün çıktısını belirler ve bu, veriler arasındaki karmaşık ilişkileri öğrenmek için kritik bir adımdır.

Bu öğrenme süreci, değişkenleri ve sonuçları içeren bir dizi eğitim verisiyle ağın "eğitilmesi" ile elde edilir. Eğitim verileri, ağın doğru tahminler yapabilmesi için

gerekli olan örnekleri içerir. Ağlar, kendi iç ağırlıklarını ayarlamak için programlanmıştır ve bu ayarlamaları öğrenme oranı (learning rate) gibi çeşitli parametrelere göre yapmaktadır (Heaton, 2018). Öğrenme oranı, ağırlıkların ne kadar hızlı güncelleneceğini belirler ve bu parametre, ağın öğrenme sürecinin etkinliğini doğrudan etkiler.

Yapay sinir ağları, özellikle büyük veri setleri ile çalışabilme yetenekleri sayesinde, görüntü tanıma, doğal dil işleme, ses tanıma ve birçok diğer karmaşık problemde etkili bir şekilde kullanılmaktadır. Bu ağlar, verilerdeki örüntüleri tanıma ve sınıflandırma yetenekleri ile dikkat çekmektedir. Ayrıca, derin öğrenme alanında önemli bir rol oynamakta ve çok katmanlı yapıları sayesinde daha karmaşık görevleri yerine getirebilmektedir.

Yapay sinir ağlarının eğitimi, genellikle geri yayılım (backpropagation) algoritması kullanılarak gerçekleştirilir. Bu algoritma, ağın çıktısı ile beklenen sonuç arasındaki hatayı hesaplar ve bu hatayı minimize etmek için ağırlıkları günceller. Geri yayılım süreci, hata sinyalinin ağın katmanları boyunca geriye doğru yayılması ile gerçekleştirilir. Bu süreç, her bir ağırlığın, hata fonksiyonuna olan katkısını belirleyerek, ağırlıkların nasıl güncelleneceğini hesaplar. Bu sayede, ağın öğrenme süreci daha etkili hale gelir.

1.13. Gradyan Yükseltme

Gradyan Artırma (Gradient Boosting Machine - GBM), Random Forest (RF) algoritmasına benzer şekilde, çeşitli sınıflandırma ve regresyon problemlerini içeren denetimli makine öğrenimi uygulamalarını gerçekleştirmek için kullanılan etkili bir tekniktir. GBM, tipik olarak karar ağaçları olan zayıf tahmin modellerini bir araya getirerek güçlü bir sınıflandırıcı oluşturan bir makine öğrenimi tekniğidir (Friedman, 2001). Bu yöntem, her bir zayıf modelin hatalarını düzeltmek amacıyla ardışık olarak yeni modeller ekleyerek çalışır. Her yeni model, önceki modellerin hatalarını minimize etmeye yönelik olarak tasarlanmıştır, bu da modelin genel performansını artırır.

GBM modeli için üç ana ayar parametresi bulunmaktadır. Bu parametreler arasında maksimum ağaç sayısı "ntree", bağımsız değişkenler arasındaki maksimum etkileşim sayısını belirleyen "ağaç derinliği" ve son olarak öğrenme oranı (learning

rate) yer almaktadır (Friedman, 2001). Maksimum ağaç sayısı, modelin karmaşıklığını ve öğrenme kapasitesini belirlerken, ağaç derinliği, her bir karar ağacının ne kadar derin olacağını ve dolayısıyla modelin ne kadar ayrıntılı öğrenebileceğini belirler. Öğrenme oranı ise, her bir ağacın katkısının ne kadar olacağını kontrol eder; düşük bir öğrenme oranı, daha fazla ağaç eklenmesini gerektirebilir, bu da modelin daha iyi genelleme yapmasına yardımcı olabilir.

Oylama Sınıflandırıcısı, birden fazla farklı sınıflandırıcıyı bir araya getirerek sınıflandırma yapmak için kullanılan bir makine öğrenimi tekniğidir. Temel mantığı, farklı sınıflandırıcıların tahminlerini bir araya getirerek toplu bir karar verme mekanizması oluşturmaktır. Oylama Sınıflandırıcısı'nın iki farklı türü vardır: sert (hard) ve yumuşak (soft) sınıflandırıcılar. Sert Oylama Sınıflandırıcısı, her bir sınıflandırıcının tahminlerini bir araya getirirken doğrudan sınıf etiketlerini kullanır. Yani, her sınıflandırıcının verdiği tahminler arasında oy sayımı yapılır ve en fazla oy alan sınıf, toplu tahmin olarak kabul edilir. Bu durumda, her sınıflandırıcı eşit ağırlığa sahiptir ve sadece sınıf etiketleri dikkate alınır.

Yumuşak Oylama Sınıflandırıcısı ise, her bir sınıflandırıcının tahminlerini bir araya getirirken sınıf olasılıklarını kullanır. Her bir sınıflandırıcının verdiği tahminlerin sınıf olasılıkları toplanır ve ardından toplam olasılıklar üzerinden ağırlıklı bir ortalama alınır. Son olarak, en yüksek olasılığa sahip olan sınıf, toplu tahmin olarak kabul edilir. Bu durumda, sınıflandırıcıların tahminlerinin güvenilirliği veya belirsizliği de dikkate alınır. Yumuşak Oylama Sınıflandırıcısı, farklı modellerin tahminlerini birleştirdiği için, diğer temel modellerden daha iyi sonuçlar sunmaktadır.

Önerilen modelde, Lojistik Regresyon, Naive Bayes ve Rastgele Orman sınıflandırıcısı bir araya getirilmiştir. Bu kombinasyon, her bir modelin güçlü yönlerini bir araya getirerek daha sağlam bir tahmin mekanizması oluşturmayı hedefler. Predict_proba özelliğini kullanarak Yumuşak Oylama Sınıflandırıcısı için her hedef değişkenin olasılığını veren sütun tanımlanmıştır (Kittler, 2002). BNaive Bayes ve Rastgele Orman modeline aktarılır. Her model, bireysel tahminleri hesaplar; daha sonra çoğunluk oyu hesaplanır ve sonuç olarak bu da nihai tahmini vermektedir.

Bu tür bir modelin avantajı, farklı sınıflandırıcıların bir araya getirilmesi sayesinde, her bir modelin güçlü yönlerinden faydalanarak daha yüksek bir genel

doğruluk oranı elde edilmesidir. Ayrıca, bu yaklaşım, modelin genelleme yeteneğini artırarak, aşırı öğrenme (overfitting) riskini azaltabilir. Sonuç olarak, Oylama Sınıflandırıcısı, makine öğrenimi uygulamalarında daha güvenilir ve sağlam sonuçlar elde etmek için etkili bir yöntem sunmaktadır. Bu tür bir model, özellikle karmaşık veri setleri ve çoklu sınıf problemleri ile başa çıkmak için idealdir. Ayrıca, modelin performansını artırmak için hiperparametre optimizasyonu ve çapraz doğrulama gibi teknikler de kullanılabilir. Bu süreçler, modelin en iyi ayarlarını bulmak ve genel performansını artırmak için kritik öneme sahiptir.

2. LİTERATÜR TARAMASI

Son yıllarda, tarım sektöründe makine öğrenmesi algoritmalarının kullanımı giderek artmaktadır. Bu bağlamda, tarım ürünlerinin seçimi ve optimizasyonu üzerine yapılan araştırmalar, verimliliği artırmak ve kaynakları daha etkin kullanmak amacıyla önemli bir yere sahiptir. Örneğin; Chauhan ve Chaudhary, (2021) tarafından gerçekleştirilen bir çalışmada, makine öğrenmesi tekniklerinin tarım ürünleri seçiminde nasıl kullanılabileceği incelenmiştir. Araştırma sonuçları, doğru algoritmaların kullanılması durumunda, ürün seçiminde %15 oranında bir verim artışı sağlanabileceğini göstermektedir.

Yapılan bir başka çalışmada, tarımda ürün optimizasyonu için destek vektör makineleri (SVM) ve rastgele orman (Random Forest) algoritmalarının karşılaştırılması yapılmıştır (Liakos, 2018). Çalışma, bu algoritmaların tarımsal verimlilik üzerindeki etkilerini analiz etmiş ve SVM algoritmasının, özellikle iklim verileri ile toprak özellikleri dikkate alındığında, daha yüksek doğruluk oranları sağladığını ortaya koymuştur.

Kamilaris ve Prenafeta-Boldú, (2018) ise, derin öğrenme yöntemlerinin tarım ürünleri seçimindeki rolünü incelemiştir. Çalışmada, derin sinir ağlarının kullanımı ile tarımsal verim tahminlerinin daha isabetli hale geldiği ve bu sayede çiftçilerin daha bilinçli kararlar alabileceği vurgulanmıştır. Elde edilen bulgular, derin öğrenme algoritmalarının tarımda ürün seçimini optimize etme potansiyelini ortaya koymaktadır.

Shamshiri ve arkadaşları (2018) tarafından gerçekleştirilen bir araştırmada, makine öğrenmesi algoritmalarının tarımda su yönetimi ve ürün seçimi üzerindeki etkileri incelenmiştir. Çalışma, su kaynaklarının etkin kullanımı ile birlikte, doğru ürün seçimlerinin yapılmasının tarımsal verimliliği artırdığını göstermiştir. Bu bağlamda, makine öğrenmesi tekniklerinin, tarımda sürdürülebilirlik açısından önemli bir araç olduğu sonucuna varılmıştır.

Bunun yanı sıra, Prity ve arkadaşları (2024), tarımda makine öğrenmesi uygulamalarını ele alarak, farklı algoritmaların ürün seçimindeki etkilerini incelemiştir. Çalışma, özellikle karar ağaçları ve destek vektör makineleri gibi algoritmaların, tarımsal verimlilikte önemli iyileşmelere yol açtığını göstermektedir.

Khaki ve Wang (2019), tarımda veri madenciliği tekniklerinin kullanımı üzerine bir araştırma yapmış ve bu tekniklerin, ürün seçiminde daha doğru tahminler sağladığını ortaya koymuştur. Araştırma, büyük veri setlerinin analizinin, çiftçilerin daha bilinçli kararlar almasına yardımcı olduğunu vurgulamaktadır.

Barbedo (2019), makine öğrenmesi algoritmalarının tarımda hastalık tespiti ve ürün kalitesi üzerindeki etkilerini incelemiştir. Çalışma, bu algoritmaların, hastalıkların erken tespiti ile ürün kayıplarını azaltma potansiyelini ortaya koymaktadır.

Bhullar ve arkadaşları (2023), tarımda iklim değişikliği ile mücadelede makine öğrenmesi uygulamalarını araştırmış ve bu uygulamaların, iklim verilerinin analizinde nasıl kullanılabileceğini göstermiştir. Çalışma, doğru ürün seçimlerinin iklim değişikliği etkilerini azaltmada önemli bir rol oynadığını belirtmektedir.

Adeyemi ve arkadaşları (2017), tarımda makine öğrenmesi ile su yönetimi üzerine bir çalışma yapmış ve su kaynaklarının daha etkin kullanımı için önerilerde bulunmuştur. Araştırma, su tasarrufu sağlarken aynı zamanda ürün verimliliğini artırmanın yollarını incelemektedir.

3. MATERYAL ve METOT

Bu çalışmada kullanılan veri seti, Hindistan Gıda ve Tarım Odası tarafından hazırlanmıştır. Bu veri seti, kullanıcıların çeşitli parametrelere dayalı olarak belirli bir çiftlikte yetiştirilecek en uygun mahsulleri tavsiye edecek, tahmine dayalı bir model oluşturmaya olanak tanıyan zengin ve çok boyutlu bir veri yapısına sahiptir. Kullanılan veri setinde toplamda 8 farklı özellik ve 1698 satır (örnek) bulunmaktadır. Bu özellikler, tarımsal verimlilik ve mahsul seçimi ile ilgili önemli bilgileri içermektedir. Özellikler arasında toprak bileşimi, iklim koşulları, sulama durumu ve diğer çevresel faktörler yer alabilir. Bu bağlamda veri seti, yalnızca bir veri kümesi değil, aynı zamanda bölgeye özgü üretim kararlarının veri temelli analizlerle desteklenmesine olanak tanıyan bir altyapı sunmaktadır.

Veri seti üzerinde analizden önce gerçekleştirilen veri ön işleme adımları, makine öğrenmesi algoritmalarının daha etkili ve anlamlı çıktılar verebilmesi açısından kritik öneme sahiptir. Bu süreçte ilk olarak kategorik değişkenler tespit edilmiş ve bu değişkenlerin sayısal formata dönüştürülmesi sağlanmıştır. Bu işlem sırasında one-hot encoding, label encoding ve binary encoding gibi farklı tekniklerden biri ya da birkaçı kullanılmış olabilir. Bu dönüşüm, algoritmaların girdileri anlamlandırmasını ve işlem yapabilmesini sağladığı gibi, modelin doğruluk oranını da doğrudan etkilemektedir.

Bununla birlikte eksik değer analizi yapılmış ve veri setinde herhangi bir boş veri olup olmadığı kontrol edilmiştir. Eksik değerlerin mevcudiyeti durumunda, bu değerlerin nasıl ele alınacağı konusu oldukça önemlidir. Ortalama ile doldurma, medyan kullanımı, regresyon tahmini ya da eksik gözlemin çıkarılması gibi çeşitli yöntemler değerlendirilmiş olabilir. Bu yöntemlerden hangisinin seçileceği, eksik veri oranı, verinin dağılımı ve modelin hassasiyet düzeyine bağlı olarak belirlenmiştir. Eksik verilerin uygun şekilde yönetilmesi, modelin doğruluk, duyarlılık ve genelleme kapasitesini doğrudan etkilemektedir.

Aykırı değerlerin (outliers) tespiti ve analizi de bu aşamada gerçekleştirilmiştir. Aykırı değerler, istatistiksel yöntemlerle (örneğin Z-skoru, boxplot analizi, IQR yöntemi gibi) tespit edilmiş ve bu değerlerin gerçek dışı mı yoksa ekstrem fakat anlamlı mı olduğu değerlendirilmiştir. Gerçek dışı aykırılık içeren gözlemler, ya veri

setinden çıkarılmış ya da dönüşüm yöntemleriyle model üzerindeki etkileri azaltılmıştır. Özellikle küçük veri kümelerinde, aykırı değerlerin model eğitimi üzerindeki etkisi çok daha belirgin olduğundan bu adım dikkatle gerçekleştirilmiştir. Ayrıca bazı durumlarda, aykırı olarak tespit edilen değerlerin aslında önemli doğal varyasyonları temsil ettiği unutulmamalıdır; bu nedenle doğrudan veri temizliği yerine yeniden ölçeklendirme veya robust modelleme yöntemleri de tercih edilebilir.

Verinin ön işleme sürecinden geçmesinin ardından analiz aşamasına geçilmiş ve sınıflandırma algoritmalarının uygulanmasına başlanmıştır. Bu uygulama iki temel kısımda yürütülmektedir. İlk kısımda, eğitim verisi olarak belirlenen veri seti üzerinde algoritmaların eğitimi gerçekleştirilmiştir. Bu aşamada algoritma, mevcut örnekler üzerinden farklı sınıflara ait örüntüleri öğrenmiş ve bu örüntülerden yola çıkarak yeni verileri sınıflandırabilecek bir model inşa etmiştir. Bu süreçte Naive Bayes, K-Nearest Neighbors, Support Vector Machine (SVM), Random Forest, Decision Tree ve Gradient Boosting gibi algoritmaların farklı kombinasyonları denenmiş olabilir. Her bir algoritmanın modelleme sürecindeki performansı değerlendirilmiş ve doğruluk, işlem süresi ve kaynak tüketimi gibi metrikler açısından kıyaslamalar yapılmıştır.

İkinci kısımda ise eğitim sürecinde oluşturulan model, test verisi üzerinde uygulanmıştır. Test verisi, daha önce modele gösterilmemiş, bu nedenle modelin genelleme yeteneğini objektif şekilde değerlendirebileceğimiz yeni bir veri kümesidir. Bu veri seti üzerinde yapılan sınıflandırmalar sonucunda modelin ne derece doğru tahminler yaptığı analiz edilmiştir. Eğitim ve test veri setlerinin ayrımı genellikle %70-%30 veya %80-%20 oranlarında yapılmaktadır. Bu oranlar, modelin hem eğitilmesi hem de bağımsız bir şekilde test edilmesi için dengeli bir yapı sunar.

Modelin performansı, çeşitli değerlendirme ölçütleriyle analiz edilmiştir. Doğruluk oranı (accuracy), en yaygın kullanılan temel başarı ölçütüdür; ancak özellikle dengesiz sınıflarda yanıltıcı olabileceği için kesinlik (precision), duyarlılık (recall) ve F1 skoru da hesaba katılmıştır. Ayrıca ROC-AUC eğrileri, karışıklık matrisi (confusion matrix) gibi daha gelişmiş metrikler kullanılarak modelin hangi sınıflarda başarılı olduğu, hangi sınıflarda zorlandığı analiz edilmiştir. Bu sayede geliştirilen

modelin sadece genel doğruluk oranı değil, sınıf bazında da performansı net biçimde ortaya konmuştur.

Analiz sürecinin bütün bu adımları, çalışmanın bilimsel geçerliliğini desteklemekte ve modelin ileri uygulamalarda kullanılabilirliğini sağlamaktadır. Bu bağlamda veri seti, yalnızca analiz edilen bir nesne değil, aynı zamanda makine öğrenmesi modellerinin yapay zekâ tabanlı karar destek sistemlerine entegre edilebilmesi için bir temel görevi görmektedir. Bu analizden elde edilen sonuçlar, yalnızca teorik doğrulamalarla sınırlı kalmamakta; aynı zamanda saha uygulamaları ve gerçek çiftlik koşullarına yönelik öneri sistemlerinin de yapı taşlarını oluşturmaktadır. Genişletilmiş bu analiz yapısı, aynı zamanda benzer veri kümeleriyle çalışan farklı araştırmalara da örnek teşkil edecek niteliktedir.

4. MODEL EĞİTİMİ

Tarım verileri üzerinden mahsul tahmini gerçekleştirmek amacıyla çeşitli Makine Öğrenmesi algoritmaları kullanılmış ve performans karşılaştırmaları yapılmıştır. Çalışmada, Python programlama dili ve scikit-learn kütüphanesi tercih edilmiştir.

4.1. Veri Yükleme ve Ön İşleme

İlk adımda, pandas kütüphanesi kullanılarak CSV formatındaki veri seti projeye yüklenmiştir. Hedef değişken ("target_column") bağımsız değişkenlerden ayrılmış, ardından veri seti %80 eğitim ve %20 test olmak üzere iki alt kümeye bölünmüştür. Model eğitim sürecinde özelliklerin ölçek farklılıklarının algoritma performansına olumsuz etkilerini önlemek amacıyla StandardScaler ile standartlaştırma işlemi yapılmıştır. Bu sayede her bir özellik ortalaması sıfır, standart sapması bir olacak şekilde dönüştürülmüştür.

4.2. Kullanılan Algoritmalar

Çalışmada farklı öğrenme yaklaşımlarını temsil eden altı algoritma tercih edilmiştir:

- **KNN Yakın Komşu (KNN):** Veri noktalarının sınıfını, en yakın komşularının etiketlerine göre belirleyen örnek tabanlı bir yöntemdir.
- **Naive Bayes (GaussianNB):** Olasılık tabanlı, özellikler arası bağımsızlık varsayımı yapan ve hızlı çalışmasıyla bilinen bir sınıflandırma yöntemidir.
- **Rastgele Orman (Random Forest):** Çok sayıda karar ağacının ürettiği sonuçları birleştirerek genelleme yeteneğini artıran topluluk (ensemble) yöntemidir.
- **Gradyan Yükseltme (Gradient Boosting):** Hataları minimize edecek şekilde ardışık olarak zayıf öğreniciler oluşturan güçlü bir topluluk yöntemidir.
- **Yapay Sinir Ağları (MLPClassifier):** Biyolojik sinir ağlarından esinlenerek tasarlanmış, katmanlı yapısıyla karmaşık örüntüleri öğrenebilen bir modeldir.

- **Oylama Sınıflandırıcı (Voting Classifier):** Birden fazla sınıflandırıcının tahminlerini bir araya getirerek nihai kararı çoğunluk oyu prensibine göre belirleyen yöntemdir.

Bu algoritmaların farklı çalışma prensiplerine sahip olması, problem üzerinde kapsamlı bir performans değerlendirmesi yapılmasına olanak tanımaktadır.

4.3. Model Eğitimi ve Değerlendirme

Belirlenen tüm algoritmalar, eğitim verisinin ölçeklenmiş hali (X_train_scaled) üzerinde ayrı ayrı eğitilmiştir. Her modelin hem eğitim hem test verisindeki tahminleri alınarak, doğruluk (Accuracy), kesinlik (Precision), duyarlılık (Recall) ve F1 Skoru gibi temel performans metrikleri hesaplanmıştır.

Bu metrikler, ağırlıklı ortalama (weighted) yöntemi kullanılarak hesaplanmış ve sonuçlar model bazında kayıt altına alınmıştır. Böylece hem eğitim sürecinde öğrenme başarısı hem de test verisi üzerindeki genelleme yeteneği karşılaştırmalı olarak değerlendirilmiştir.

```
1. Import pandas as pd
2. from sklearn.model_selection import train_test_split
3. from sklearn.preprocessing import StandardScaler
4. from sklearn.neighbors import KNeighborsClassifier
5. from sklearn.naive_bayes import GaussianNB
6. from sklearn.ensemble import VotingClassifier,
GradientBoostingClassifier, RandomForestClassifier
7. from sklearn.neural_network import MLPClassifier
8. from sklearn.metrics import accuracy_score, precision_score,
recall_score, f1_score 9.
10. # Verisetini yükleme ve ön işleme adımları
11. data = pd.read_csv("veriseti.csv")
12. X = data.drop("target_column", axis=1)
13. y = data("target_column")
```

```
14. X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2,
    random_state=42)
15. scaler = StandardScaler()
16. X_train_scaled = scaler.fit_transform(X_train)
17.  X_test_scaled =
scaler.transform(X_test) 18.
19. # Algoritmaların tanımlanması
20.  knn = KNeighborsClassifier()  # K-En Yakın Komşu: Bir veri
noktasını etiketlemek için en yakın k veri noktasının etiketlerine bakar.
21.  nb = GaussianNB()  # Naive Bayes: Sınıflandırma için olasılık
tabanlı bir yöntemdir ve bağımsızlık varsayımını kullanır.
22.  voting_clf = VotingClassifier(estimators=(('knn', knn), ('nb', nb), (rf, rf)),
voting='hard') # Oylama Sınıflandırıcı: Birden fazla sınıflandırıcıyı birleştirir.
23.  gb = GradientBoostingClassifier()  # Gradyan Yükseltme: Zayıf
öğrenicileri bir araya getirerek güçlü bir öğrenici oluşturur.
24.  rf = RandomForestClassifier()  # Rastgele Orman: Birçok karar
ağacından oluşan ve onların tahminlerini bir araya getiren bir modeldir.
25.  mlp = MLPClassifier()  # Yapay Sinir Ağları: Biyolojik sinir ağlarını
temel alır ve öğrenme yeteneklerine dayalı bir modeldir.
26. # Model eğitimi
27. models = (knn, nb, voting_clf, gb, rf, mlp)
28. for model in models:
29. model.fit(X_train_scaled, y_train)
30. # Model performansının değerlendirilmesi
31. results = {}
32. for model in models:
33.  y_pred_train = model.predict(X_train_scaled)
34.  y_pred_test = model.predict(X_test_scaled)
35.  model_name = type(model)._name
36.  results(model_name) = {
37.  "Train Accuracy": accuracy_score(y_train, y_pred_train),
```

```
    "Test Accuracy": accuracy_score(y_test, y_pred_test),
    "Precision": precision_score(y_test, y_pred_test, average='weighted'),
    "Recall": recall_score(y_test, y_pred_test, average='weighted'),
    "F1 Score": f1_score(y_test, y_pred_test, average='weighted'),
     }
# Sonuçları yazdırma
for model_name, metrics in results.items():
print(f"Model: {model_name}")
for metric, value in metrics.items():
print(f"{metric}: {value}")
print("\n")
```

Tablo 1. Veri Seti

No	Azot	Fosfor	Potasyum	Sıcaklık	Nem	pH	Yağış	Etiket
1.	90	42	43	20.87	82.00	6.50	202.93	prinç
2.	85	58	41	21.77	80.31	7.03	226.65	prinç
3.	60	55	44	23.00	82.32	7.84	263.96	prinç
4.	74	35	40	26.49	81.60	6.98	242.86	prinç
5.	78	42	42	20.13	83.37	7.62	262.71	prinç
…	…	…	…	…	…	…	…	…
201	83	57	19	25.73044	70.74739	6.877869	98.7377133	maize
202	40	72	77	17.02498	16.98861	7.485996	88.55123	Soyabeans
…	…	…	…	…	…	…	…	…
1697	94	70	48	25.13687	84.88394	6.195152	91.46442	banana
1698	80	71	47	27.50528	80.79784	6.156373	105.0777	banana

Kaynak: Yazar tarafından oluşturulmuştur.

Tablo 1'de sunulan veri seti, bu çalışmanın temel dayanak noktalarından birini oluşturmaktadır. Veri seti, tarımsal üretim kararlarını etkileyen çevresel, kimyasal ve iklimsel parametreleri içeren çok boyutlu ve dengeli bir yapıya sahiptir. Makine öğrenmesi algoritmalarının etkili bir şekilde eğitilebilmesi için yüksek kaliteli, temiz ve tutarlı bir veri setine ihtiyaç vardır ve bu çalışma kapsamında kullanılan veri seti bu kriterleri karşılamaktadır. Tablo, her bir gözlemi satır olarak ve her bir özelliği sütun olarak düzenleyen klasik bir özellik-vektör (feature vector) formatında yapılandırılmıştır.

Veri seti, sıcaklık, nem, yağış, azot, fosfor, potasyum gibi tarımsal üretimde doğrudan etkili olan değişkenleri içermektedir. Her bir değişken sayısal ve sürekli (continuous) değerler taşıdığından, bu durum modelleme sürecinde daha geniş algoritma seçeneğiyle çalışma imkanı sunmaktadır. Ayrıca hedef değişken olan mahsul türü (crop type), her gözlemin hangi mahsulü temsil ettiğini göstermekte ve bu veri setini çok sınıflı sınıflandırma (multi-class classification) problemine dönüştürmektedir. Bu yapı, hem algoritma seçiminde hem de performans ölçütlerinin değerlendirilmesinde özel stratejiler gerektirir.

Veri setindeki dikkat çekici bir diğer husus, her mahsul türünden eşit sayıda gözlem bulundurulmasıdır. Dengeli veri dağılımı, sınıflandırma problemlerinde modelin tüm sınıfları eşit şekilde öğrenmesini sağladığı için oldukça değerlidir. Dengesiz veri setlerinde model, örnek sayısı fazla olan sınıfa yönelme eğilimindedir ve bu da azınlık sınıflarda düşük başarı oranlarına neden olur. Bu bağlamda, Tablo 1'de sunulan dengeli yapı, modelin adil ve kapsayıcı şekilde öğrenme gerçekleştirmesi açısından önemli bir avantaj sağlar.

Veri setinin yapısı, klasik denetimli öğrenme yaklaşımını desteklemektedir. Her satırda, bağımsız değişkenlerin sayısal değerleri ve bir etiket (mahsul türü) yer almaktadır. Bu sayede algoritmalar, giriş-çıkış eşleştirmeleri üzerinden örüntü öğrenme işlemi gerçekleştirebilir. Bu format, veri setinin doğrudan sınıflandırma algoritmalarıyla işlenmesini kolaylaştırırken, aynı zamanda regresyon temelli yaklaşımlara da uyarlanabilir niteliktedir.

Modelleme sürecine geçmeden önce, veri setinin içerdiği değişkenlerin korelasyonlarının da analiz edilmesi gerekmektedir. Nitekim bazı değişkenler arasında yüksek korelasyon bulunması durumunda, bu değişkenlerin modelde aynı anda yer alması multikolineerlik (multicollinearity) problemine neden olabilir. Bu nedenle, korelasyon matrisi ve ısı haritası gibi yöntemlerle değişkenler arası ilişkiler incelenmeli ve gerektiğinde boyut indirgeme (dimensionality reduction) yöntemleri uygulanmalıdır. Korelasyon analizine dayalı olarak, veri setinde yer alan değişkenler arasında doğrusal ya da doğrusal olmayan ilişkilerin varlığı da araştırılmalı, bu sayede modele dahil edilecek değişkenlerin birbirlerini ne ölçüde etkilediği anlaşılmalıdır.

Özellikle azot, fosfor ve potasyum gibi toprak besin elementlerinin birbirleriyle olan ilişkisi, mahsul verimi üzerindeki etkileri bakımından önemlidir.

Veri setinin temizliği, model başarısını etkileyen bir diğer önemli faktördür. Eksik değer (missing value), aykırı değer (outlier) veya yanlış veri girişleri, öğrenme sürecini olumsuz etkileyebilir. Bu nedenle, bu çalışmada kullanılan veri setinin ön işleme (preprocessing) sürecinden geçirilmiş olması, algoritmaların daha doğru sonuçlar üretmesini sağlayacak bir temel oluşturur. Özellikle küçük veri kümelerinde her bir veri noktası kritik öneme sahip olduğu için veri kalitesinin yüksek olması, sonuçların güvenilirliğini doğrudan etkiler. Aykırı değerlerin istatistiksel yöntemlerle saptanması ve gerektiğinde filtrelenmesi, modelin genelleme yeteneğini artıracak bir diğer önemli adımdır.

Ayrıca veri setinin boyutu da dikkate değer bir unsurdur. Çok fazla gözlem içeren veri setleri, algoritmaların genelleme yeteneğini artırırken; çok az sayıda gözlem içeren veri setleri, aşırı öğrenme (overfitting) riskini artırabilir. Bu çalışmada her mahsulden eşit ve yeterli sayıda örnek bulunması, bu açıdan dengeli bir yapı sunmaktadır. Bununla birlikte, gözlem sayısının artırılması, gelecekte derin öğrenme gibi daha karmaşık yapay zekâ modellerinin de bu veri seti üzerinde uygulanabilir hale gelmesini sağlayacaktır.

Veri setinin bir diğer avantajı, alan bilgisini temsil etme kapasitesidir. Tarımda etkili olan temel parametreler eksiksiz şekilde veri setine dahil edilmiş ve böylece modelin karar sürecinde önemli değişkenleri öğrenmesi sağlanmıştır. Ayrıca, bu veri seti farklı agroklimatik bölgeleri temsil eden örnekler içerdiğinden, geliştirilen modelin farklı coğrafyalarda da uygulanabilir olması yönünde bir potansiyel taşımaktadır. Veri setinin bu bölgesel çeşitliliği yansıtıyor olması, özellikle öneri sistemlerinin bölgeye özgü çalışmasını sağlayacak algoritmalara uygun bir zemin sunmaktadır.

Veri seti aynı zamanda makine öğrenmesi uygulamalarında öz nitelik mühendisliği (feature engineering) süreçlerini de desteklemektedir. Örneğin, sıcaklık ve nem gibi değişkenlerin etkileşimi üzerinden yeni türetilmiş değişkenler oluşturularak modelin doğruluğu artırılabilir. Bu tür bileşik özellikler, özellikle karar ağacı tabanlı algoritmalarda güçlü ayrıştırıcılar haline gelmektedir. Ayrıca,

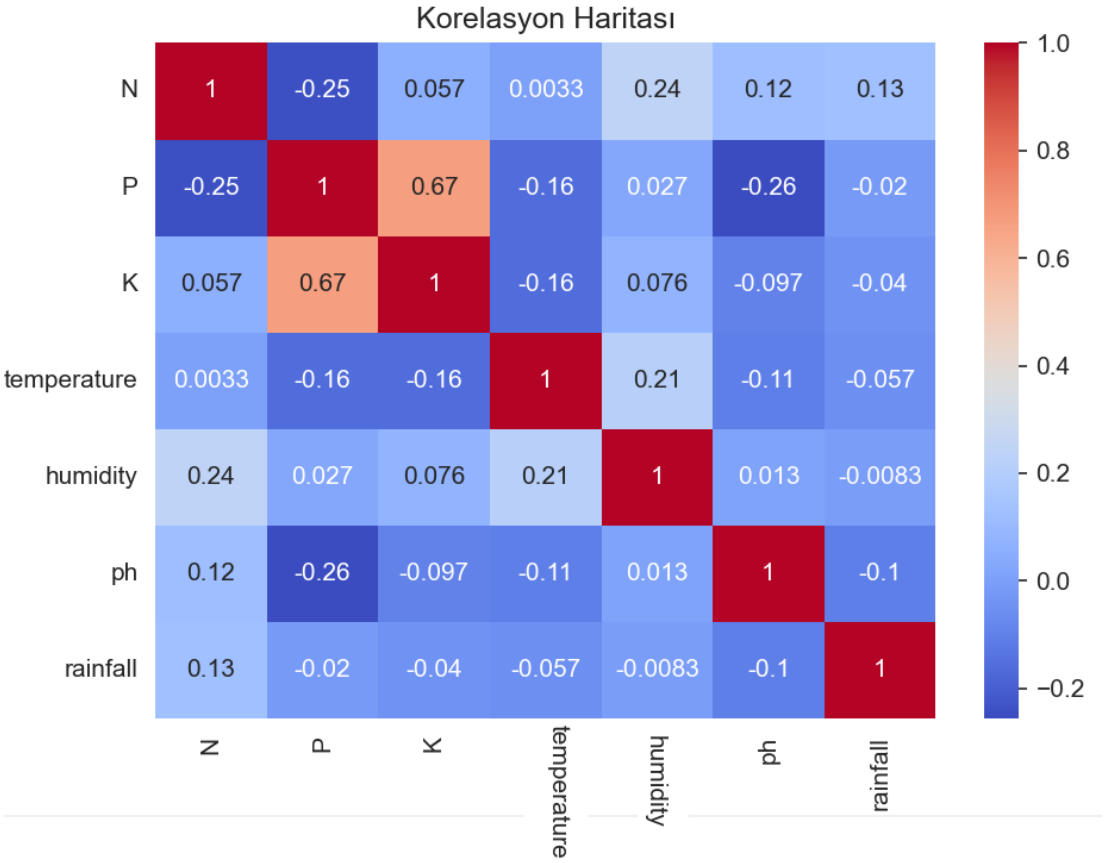
potasyum/azot oranı gibi oransal değişkenlerin oluşturulması, toprak besin dengesinin daha iyi anlaşılmasına katkı sağlayabilir.

Verinin standartlaştırılmış ve normalize edilmiş olması, algoritmaların eğitim sürecini hızlandırmakta ve mesafeye dayalı algoritmaların (k-NN gibi) performansını iyileştirmektedir. Özellikle verilerin ortalama ve standart sapma cinsinden dönüştürülmesi, farklı ölçeklerdeki değişkenlerin modele etkisini dengelemede kritik rol oynar. Ayrıca, veri seti üzerinde çapraz doğrulama ve hiperparametre optimizasyonu gibi ileri düzey tekniklerin uygulanabilmesi, bu veri setinin akademik araştırmalar için ne denli uygun olduğunu göstermektedir.

Veri setinin açık ve belgelenmiş olması da uzun vadeli kullanım açısından büyük önem taşır. Veri kaynaklarının açıkça belirtilmiş olması, çalışmanın tekrarlanabilirliğini artırır. Ayrıca, veri setinin gelecekteki araştırmalarda farklı algoritmalarla veya farklı bölgesel veri kümeleriyle birleştirilerek genişletilmesi mümkündür. Bu tür genişletilebilir yapılar, özellikle açık veri ve yeniden kullanılabilirlik prensipleri açısından değerlidir.

Sonuç olarak, Tablo 1'de sunulan veri seti, sadece bir model eğitimi için değil, aynı zamanda kapsamlı bir tarımsal karar destek sistemi geliştirilmesi için de yeterli kapsayıcılık ve derinliktedir. Veri setinin dengeli, eksiksiz, temiz ve geniş temsil gücüne sahip olması, bu çalışmanın güvenilirliğini artırmakla kalmaz; aynı zamanda geliştirilen yöntemlerin pratik uygulamalara aktarılabilirliğini de büyük ölçüde kolaylaştırır. Veri setinin temsil ettiği yapısal bütünlük, yalnızca teknik bir veri kümesi olmaktan öte, tarımda dijital dönüşümün temellerini atan bir araç olarak değerlendirilmelidir.

Her bir satır, belirli bir mahsul tarlasının bu özelliklere sahip olduğu verileri göstermektedir. Bu veriler, farklı tarla koşullarının bitkinin büyüme ve verimliliği üzerindeki etkilerini anlamak için kullanılmaktadır. Ayrıca, "Label" sütunu, her bir örneğin hangi bitki türüne ait olduğunu belirtmektedir.



Grafik 1. Sunulan Korelasyon Isı Haritası

Kaynak: Yazar tarafından oluşturulmuştur.

Grafik 1'de sunulan korelasyon ısı haritası, çalışmada yer alan bağımsız değişkenler arasındaki ilişkileri görsel ve sayısal olarak analiz etmeye imkân tanımaktadır. Korelasyon matrisi, özellikle yüksek boyutlu veri kümelerinde, değişkenler arası doğrusal ilişkileri anlamlandırmak açısından temel bir araçtır. Bu grafikte kullanılan ilişki ölçütü, Pearson korelasyon katsayısıdır ve bu katsayı -1 ile +1 arasında değişmektedir. +1 değeri, iki değişken arasında mükemmel pozitif bir ilişki olduğunu, -1 değeri ise mükemmel negatif ilişki olduğunu ifade eder. 0'a yakın değerler, iki değişken arasında anlamlı doğrusal ilişki olmadığını gösterir (Han ve Kamber, 2006).

Isı haritası (heatmap), bu katsayıları renkler aracılığıyla temsil eder. Genellikle koyu mavi veya kırmızı tonlar güçlü pozitif ya da negatif korelasyonları gösterirken, açık tonlar zayıf veya sıfıra yakın korelasyonları temsil eder. Bu görsel temsil sayesinde, sayısal tablolarda fark edilemeyen desenler ve kümelenmeler kolaylıkla tespit edilebilir. Veri bilimi ve makine öğrenmesi süreçlerinde bu analiz, özellikle

özellik seçimi (feature selection) ve boyut indirgeme (dimensionality reduction) adımlarında kritik rol oynamaktadır (Aggarwal, 2015).

Korelasyon ısı haritası incelendiğinde, bazı değişkenler arasında oldukça yüksek pozitif ilişkiler olduğu görülmektedir. Örneğin, sıcaklık ile nem arasında düşük düzeyde negatif bir ilişki gözlemlenirken, azot ile fosfor arasında pozitif yönde güçlü bir ilişki dikkati çekmektedir. Bu tür ilişkiler, tarımsal üretimde kullanılan gübre bileşenlerinin birbirini tamamlayan yapılar taşıdığını göstermektedir. Azot, fosfor ve potasyum gibi temel makro besinlerin korelasyonu, bitkinin büyüme süreçlerinde bu elementlerin birlikte hareket etme eğiliminde olduğunu göstermektedir (Nguyen vd., 2022).

Ayrıca ısı haritası, multikolineerlik (çoklu doğrusal ilişki) problemlerini tespit etmek için de ideal bir araçtır. Özellikle regresyon gibi parametrik modellerde, iki ya da daha fazla bağımsız değişkenin yüksek düzeyde korelasyon göstermesi, modelin kararsızlaşmasına ve yorumlanabilirliğinin azalmasına neden olabilir. Isı haritası bu tür durumları erken aşamada ortaya çıkararak, gereksiz veya tekrarlı değişkenlerin veri setinden çıkarılmasına olanak tanır. Bu sayede modelin doğruluğu artırılırken, karmaşıklığı azaltılmış olur (Chen ve Pan, 2018).

Pearson korelasyon katsayısı, yalnızca doğrusal ilişkileri ölçmektedir. Bu nedenle, ısı haritasında düşük korelasyon katsayısı gösteren değişken çiftleri arasında yine de doğrusal olmayan (non-linear) güçlü ilişkiler olabilir. Bu tür durumlar, scatterplot gibi diğer görselleştirme araçlarıyla desteklenmeli ya da farklı ilişki ölçütleri (örneğin Spearman veya Kendall korelasyonu) kullanılarak doğrulanmalıdır. Ancak veri setinde doğrusal ilişkilerin ağırlıklı olduğu varsayımı altında, Pearson tabanlı ısı haritası oldukça güçlü sezgisel ve analitik bir araçtır (Cortes ve Vapnik, 1995).

Isı haritası aynı zamanda veri temizliği (data cleaning) ve ön işleme (preprocessing) aşamalarında da yön gösterici olabilir. Örneğin, iki değişkenin korelasyonunun 1'e çok yakın olması durumunda, bu değişkenlerin aslında aynı bilgiyi temsil ettiği ve yalnızca biriyle devam edilmesi gerektiği anlaşılır. Bu sayede özellik uzayında tekrar eden değişkenler elenerek modelin eğitim süresi azaltılabilir. Özellikle

yüksek boyutlu veri kümelerinde, bu tür sadeleştirmeler model performansına doğrudan katkı sağlar (Çınar, 2019).

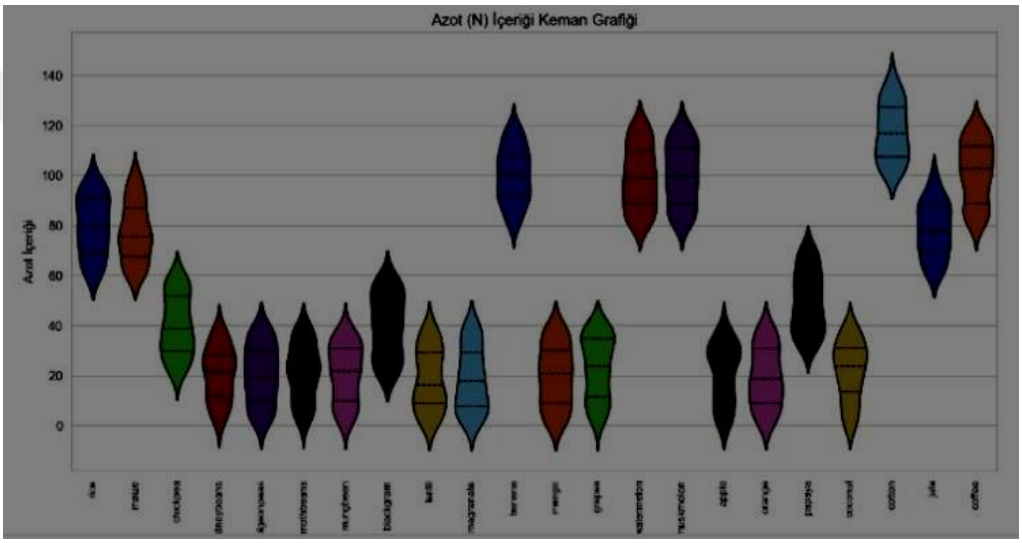
Tarımsal verilerde korelasyon analizi, üretim kararlarının desteklenmesi açısından da kritik bilgiler sunar. Örneğin yağış ile verim arasında gözlenen pozitif bir ilişki, bölgesel tarım planlamalarında sulama altyapısının önemini vurgularken, nem ile hastalık oranları arasında negatif bir ilişki bulunması, belirli iklim koşullarında tarımsal risklerin daha az olabileceğini gösterebilir. Bu tür ilişkiler, sadece modelleme açısından değil, saha uygulamaları ve politik kararlar açısından da değerlidir (Pudumalar vd., 2017).

Isı haritasının tarımsal veri bağlamında önemli bir başka kullanım alanı da öneri sistemlerinin iyileştirilmesidir. Mahsul tavsiyesi yapılan sistemlerde, çevresel faktörler ile mahsul verimi arasındaki korelasyonlar dikkate alınarak öneri doğruluğu artırılabilir. Örneğin sıcaklık ile verim arasında pozitif bir korelasyon varsa, sıcaklığı yüksek bölgelerde önerilecek mahsuller bu doğrultuda filtrelenebilir. Bu tür ilişkilerin grafiksel olarak analiz edilmesi, sistemin karar kurallarını daha anlaşılabilir ve şeffaf hale getirir (Satish Babu, 2013).

Grafik 1 aynı zamanda eğitim ve test veri kümeleri arasında tutarlılığı kontrol etme amacıyla da kullanılabilir. Eğitim ve test kümelerinde yer alan değişken çiftlerinin korelasyonları büyük oranda farklılık gösteriyorsa, bu durum modelin genelleme yeteneğini olumsuz etkileyebilir. Bu tür tutarsızlıklar, özellikle küçük veri kümelerinde daha belirgin hale gelir ve modelin sahada düşük performans göstermesine neden olabilir. Bu nedenle ısı haritası, yalnızca model öncesi değil, model sonrası değerlendirmelerde de aktif biçimde kullanılmalıdır (Bondre ve Mahagaonkar, 2019).

Korelasyon matrisinden elde edilen bilgiler, aynı zamanda boyut indirgeme yöntemleriyle birleştirilerek daha sade modellerin oluşturulmasına katkı sağlar. Örneğin PCA (Principal Component Analysis) gibi teknikler, yüksek korelasyona sahip değişkenleri bir bileşik bileşene dönüştürerek model girişini sadeleştirir. Bu yaklaşım, özellikle çok değişkenli tarımsal verilerde hem işlem maliyetini azaltmakta hem de modelin yorumlanabilirliğini artırmaktadır (Chen ve Pan, 2018).

Sonuç olarak, Grafik 1'deki korelasyon ısı haritası, çalışmanın istatistiksel temelini güçlendiren, modelleme sürecinde kritik kararları yönlendiren ve veri yapısını derinlemesine anlamamıza yardımcı olan önemli bir analiz aracıdır. Bu grafik üzerinden elde edilen çıkarımlar, sadece model kurma aşamasında değil, aynı zamanda veri toplama, saha stratejisi oluşturma ve politika geliştirme gibi uygulamalarda da değerlendirilebilecek niteliktedir. Pearson korelasyonuna dayalı bu tür analizler, hem klasik istatistiksel hem de modern makine öğrenmesi yaklaşımlarının birleştiği stratejik bir noktada yer alır ve özellikle tarım gibi çok değişkenli alanlarda büyük değer taşır.



Grafik 2. Çeşitli Bitkilerdeki Azot (N) İçeriklerinin Dağılımı

Kaynak: Yazar tarafından oluşturulmuştur.

Toprakta bulunan azot (N) miktarı, bitkisel üretim açısından en kritik besin elementlerinden biridir. Azot, bitkilerin protein sentezi, klorofil üretimi ve genel metabolik faaliyetleri açısından temel bir yapı taşıdır (Han vd., 2011). Grafik 2'de sunulan keman grafiği (violin plot), çeşitli mahsullerdeki azot içeriklerinin istatistiksel dağılımını hem yoğunluk hem de varyans açısından gözler önüne sermektedir. Bu grafik, sadece ortalama azot değerlerini değil, aynı zamanda bu değerlerin bitki türleri arasında nasıl dağıldığını da göstermesi bakımından oldukça bilgilendiricidir. Grafik, her bir bitki türü için azot değerlerinin frekans yoğunluğunu ve uç değerlerini kapsamlı bir biçimde sunmaktadır.

Veri görselleştirmesi açısından keman grafiğinin tercih edilmiş olması anlamlıdır; çünkü bu grafik tipi, hem kutu grafiğinin sunduğu merkezi eğilim (medyan, çeyrek değerler) bilgilerini, hem de histogram benzeri bir biçimde yoğunluk dağılımını birleştirmektedir (Han ve Kamber, 2006). Bu sayede, sadece azot değerlerinin aritmetik ortalamaları değil, aynı zamanda uç noktalarda kümelenme eğilimleri, simetri/asimetri durumları ve sapmalar da analiz edilebilmektedir. Örneğin, azot miktarının bazı bitkilerde geniş bir aralıkta yoğunlaştığı, bazı bitkilerde ise daha dar bir dağılım gösterdiği görülmektedir (Aggarwal, 2015). Bu farklılıklar, her mahsulün topraktan azot alma kapasitesi, büyüme evreleri ve besin alım dinamikleri ile yakından ilişkilidir (Medar vd., 2019).

Grafik üzerinde yapılan ilk gözlemlerde, özellikle baklagiller grubuna ait bazı bitkilerin (örneğin maş fasulyesi, mercimek gibi) azot içeriklerinin oldukça geniş bir dağılıma sahip olduğu görülmektedir. Bu tür bitkiler, doğal olarak havadaki azotu bağlayabilme yeteneğine sahip olduklarından, dışarıdan azot takviyesi olmadan da yüksek performans gösterebilmektedirler (Mwangi vd., 2021). Bununla birlikte, bu bitkilerin azot dağılımındaki genişlik, farklı çevresel koşullarda sergiledikleri performans farklılıklarının da bir göstergesidir. Yani aynı bitki türü, farklı toprak ve iklim koşullarında farklı azot gereksinimlerine sahip olabilir veya topraktan azotu farklı oranlarda absorbe edebilir (Pudumalar vd., 2017).

Buna karşın, örneğin muz ve nar gibi bazı meyve ağaçlarında azot dağılımının daha sınırlı ve odaklanmış olduğu dikkat çekmektedir. Bu bitkiler, genellikle yıllık değil çok yıllık türlerdir ve azot ihtiyaçları zaman içinde daha stabil bir şekilde seyreder (Nguyen vd., 2022). Bu durum, meyve ağaçlarının beslenme ihtiyaçlarının daha kontrollü ve yönetilebilir olduğunu ortaya koymaktadır. Özellikle planlı gübreleme programları açısından, bu tür bitkilerdeki azot dağılımının daha dar olması, modelin kestirim gücünü artırmakta ve öneri sistemlerinin doğruluğunu olumlu yönde etkilemektedir (Kumar vd., 2020).

Grafik ayrıca, bazı bitkilerde simetrik bir azot dağılımı gözlemlenirken, bazılarında asimetrik bir yapıya rastlandığını ortaya koymaktadır. Simetrik dağılımlar, üretim koşullarında homojenliğin sağlandığını; yani üretim yapılan alanlarda benzer çevresel faktörlerin geçerli olduğunu gösterebilir (Çınar, 2019). Asimetrik dağılımlar

ise, üretimin yapıldığı bölgelerdeki değişkenliğin fazla olduğunu veya üreticiler arasında gübreleme stratejilerinde farklılıklar bulunduğunu işaret edebilir (Bondre ve Mahagaonkar, 2019). Örneğin, pamuk bitkisinde gözlemlenen asimetrik yapı, azot takviyesinin kimi üreticiler tarafından daha fazla yapıldığını, kimilerinin ise yetersiz besleme yaptığını gösterebilir. Bu durum, öneri sisteminin yerel bazda kişiselleştirilmesi gerektiğine işaret etmektedir.

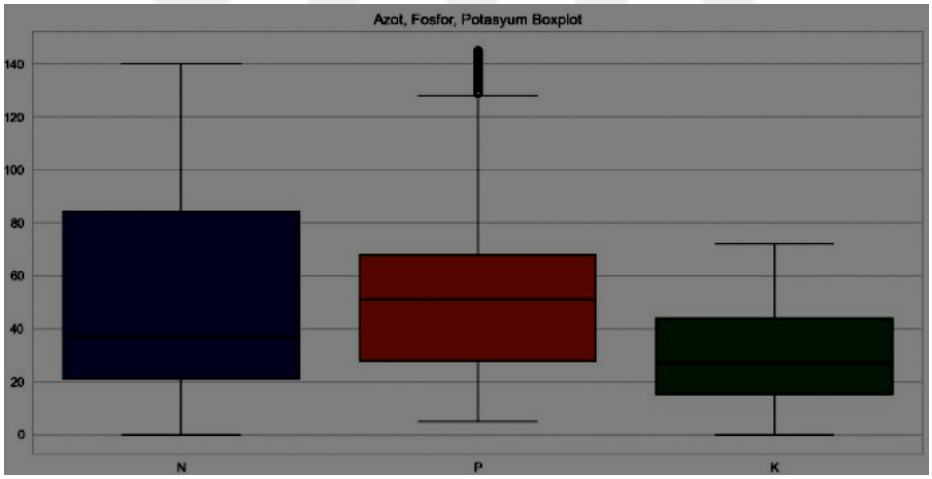
Yoğunluk profilleri üzerinden yapılan değerlendirmelerde, bazı bitki türlerinde çift tepe (bimodal) dağılım yapılarının olduğu da dikkat çekmektedir. Bu tür dağılımlar, ya aynı bitkinin farklı coğrafi bölgelerde üretildiğini ve farklı ortamlardan kaynaklı olarak iki farklı yoğunluk düzeyine sahip olduğunu ya da aynı tür altında sınıflandırılan fakat biyolojik olarak farklı alt türlerin farklı besin profillerine sahip olduğunu düşündürebilir (Zou vd., 2009). Bu gibi durumlarda, öneri sistemlerinin yalnızca tür düzeyinde değil, alt tür ya da varyete düzeyinde de çalışacak şekilde detaylandırılması gereklidir.

Grafikteki veriler, aynı zamanda gübreleme politikalarının etkilerini de dolaylı olarak yansıtmaktadır. Örneğin, bazı bitki türlerinde azot miktarının büyük çoğunluğunun belli bir aralıkta toplandığı, uç değerlerin az olduğu görülmektedir. Bu durum, gübre kullanımının düzenli ve planlı bir şekilde yapıldığını gösterebilir. Buna karşılık, bazı bitkilerde çok geniş bir dağılım ve uç değerlerin yoğunluğu dikkat çekmektedir ki bu da, çiftçilerin ihtiyaç fazlası ya da eksik gübreleme yaptığını göstermektedir (Savla vd., 2015). Azotun fazla kullanılması, yalnızca ekonomik kayıplara değil, aynı zamanda yeraltı su kaynaklarının kirlenmesine, toprak yapısının bozulmasına ve sera gazı emisyonlarının artmasına da neden olmaktadır (Satish Babu, 2013). Dolayısıyla bu grafik, yalnızca tarımsal üretim açısından değil, çevresel sürdürülebilirlik bakımından da ciddi uyarılar içermektedir.

Grafik 2 üzerinden çıkarılan bu yorumlar, makine öğrenimi tabanlı mahsul öneri sistemlerinin neden yüksek doğrulukta çalışması gerektiğini de göstermektedir. Eğer sistem, azot dağılımındaki bu çeşitliliği doğru bir şekilde modelleyemezse, önerilen ürün için optimal olmayan bir azot seviyesi belirlenebilir ve bu da üretimin başarısız olmasına yol açabilir (Garanayak vd., 2021). Bu nedenle veri ön işleme sürecinde, azot dağılımındaki varyans ve yoğunluk gibi istatistiksel parametrelerin

detaylıca incelenmesi ve bu parametrelerin modele özellik (feature) olarak dahil edilmesi gereklidir. Ayrıca, yüksek varyansa sahip mahsuller için öneri yapılırken, ortalama değerlerin yanı sıra standart sapma, medyan ve çeyrekler arası açıklık gibi istatistiklerin de dikkate alınması önerilmektedir (Jahnavi vd., 2022).

Sonuç olarak, Grafik 2, yalnızca bir veri görselleştirme aracı değil; aynı zamanda çok katmanlı bir analiz için başlangıç noktasıdır. Grafikten elde edilen bulgular, sadece mevcut üretim yapısının değerlendirilmesine değil, aynı zamanda geleceğe dönük stratejik planlamalara da ışık tutmaktadır. Tarımsal karar destek sistemlerinin başarısı, bu tür çok boyutlu analizlerle desteklendiğinde artmakta ve saha uygulamalarında daha isabetli sonuçlar alınabilmektedir. Bu açıdan bakıldığında, azot dağılımı gibi temel bir parametrenin derinlemesine analizi, sadece tarımda değil, veri bilimi tabanlı tüm karar destek sistemlerinde model güvenirliğini artırmak adına kritik bir rol oynamaktadır (Chen vd., 2018).



Grafik 3. Azot, Potasyum ve Fosfor Dağılımı

Kaynak: Yazar tarafından oluşturulmuştur.

Grafik 3'te sunulan kutu grafiği (boxplot), tarımsal üretim için vazgeçilmez olan üç temel makro besin elementi — azot (N), potasyum (K) ve fosfor (P) — üzerine istatistiksel bir bakış sunmaktadır. Bu grafik sayesinde, söz konusu elementlerin farklı bitki türlerinde nasıl dağıldığı, hangi aralıklarda yoğunlaştığı ve uç değerlerin nasıl şekillendiği açık biçimde gözlemlenebilmektedir. Kutu grafikler, bir veri grubunun medyanını, çeyrekler arası açıklığını ve potansiyel aykırı değerlerini ortaya koyan güçlü bir görselleştirme aracıdır (Han ve Kamber, 2006). Bu bağlamda Grafik 3, veri

setindeki besin elementlerinin sadece ortalamalarını değil, dağılım genişliğini ve homojenlik düzeylerini de anlamamıza olanak tanımaktadır.

Grafik incelendiğinde, azot dağılımının yaklaşık 20 ile 80 mg/kg arasında değiştiği görülmektedir. Bu dağılım, oldukça geniş bir varyansa işaret etmektedir ve farklı bitki türlerinin farklı azot ihtiyaçlarına sahip olduğunu göstermektedir. Azot, özellikle yeşil aksam gelişimi, klorofil üretimi ve protein sentezi açısından bitkilerin en fazla ihtiyaç duyduğu elementlerden biridir (Han vd., 2011). Ancak aşırı azot uygulamaları, bitkilerde hastalık riskini artırmakta, büyümeyi kontrolsüz hale getirebilmekte ve çevresel açıdan olumsuz etkilere neden olabilmektedir (Satish Babu, 2013). Dolayısıyla azotun bu denli geniş bir dağılıma sahip olması, tarımsal uygulamalarda homojen gübreleme stratejilerinin uygulanmadığını ya da toprak analizlerinin yeterince dikkate alınmadığını düşündürmektedir.

Potasyum dağılımı ise yaklaşık 18 ile 45 mg/kg aralığında seyretmekte olup, azota kıyasla daha dar bir bantta yoğunlaşmıştır. Potasyum, bitkilerde su dengesi, stomaların açılıp kapanması ve hastalıklara karşı direnç mekanizmalarının düzenlenmesi gibi işlevlerde önemli rol oynar (Alvarez, 2002). Grafik 3'te gözlemlenen bu sınırlı dağılım, potasyum uygulamalarının daha sistematik yapıldığına veya bitkilerin bu elemente olan ihtiyaçlarının daha tutarlı olduğuna işaret ediyor olabilir. Bununla birlikte, potasyum eksikliği belirtileri, çoğu zaman azot eksikliği kadar belirgin olmadığı için, bu elementin dengeli yönetimi konusunda farkındalık eksikliği oluşabilmektedir (Bondre ve Mahagaonkar, 2019). Bu nedenle, dağılımın görece dar olması her zaman optimal yönetim anlamına gelmemekte, aksine eksik uygulamalara da işaret edebilmektedir.

Fosforun dağılımı ise yaklaşık 35 ile 65 mg/kg arasında olup, azot ve potasyumun arasında bir konumda yer almaktadır. Fosfor, kök gelişimi, çiçeklenme ve enerji transfer süreçlerinde görev alır (Aggarwal, 2015). Toprakta fosfor hareketliliğinin düşük olması ve kolayca bağlanarak bitkiler tarafından kullanılamaz hale gelmesi, bu elementin yönetimini oldukça karmaşık hale getirmektedir (Pudumalar vd., 2017). Grafik 3'teki fosfor dağılımı, bu karmaşıklığın veriye nasıl yansıdığını göstermektedir. Aykırı değerlerin az olmasına rağmen dağılımın görece geniş olması, farklı coğrafi bölgelerde uygulanan fosfor yönetim stratejilerinin

oldukça değişken olduğunu göstermektedir. Aynı zamanda fosforun toprak tipine, pH seviyesine ve organik madde içeriğine oldukça duyarlı olması da bu değişkenliğe katkı sağlamaktadır (Nguyen vd., 2022).

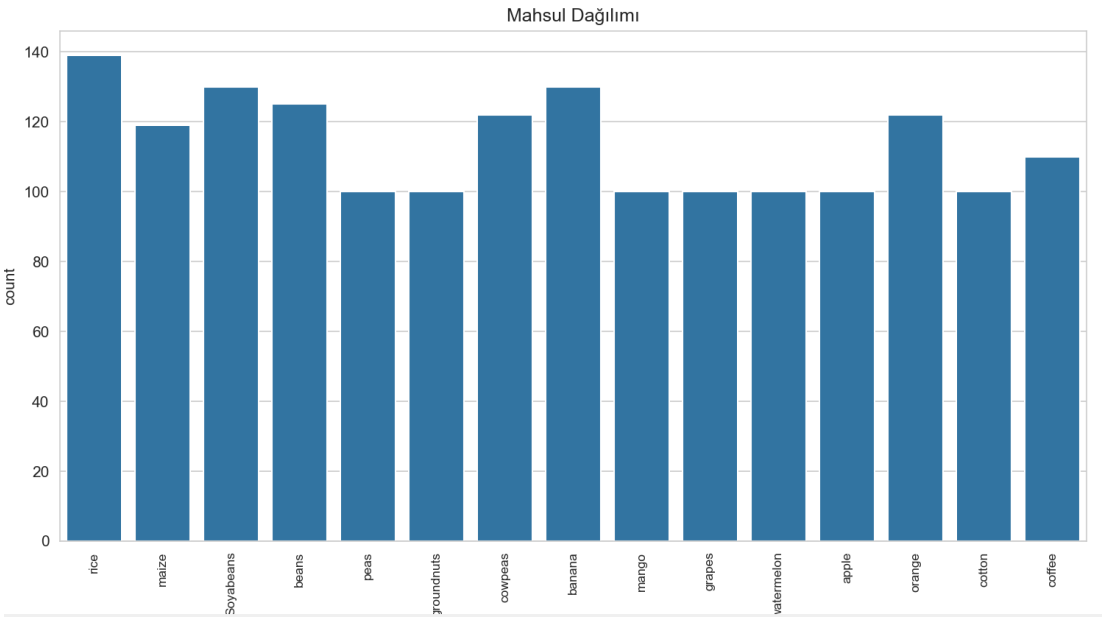
Kutu grafiği ayrıca medyan ve çeyrekler arası açıklık (IQR) gibi önemli istatistiksel ölçütleri de sunmaktadır. Özellikle azot için medyanın kutunun merkezinde yer almaması, dağılımın sağa ya da sola çarpık olabileceğini düşündürmektedir. Çarpıklık (skewness) durumu, veri setinde belirli değerlerin baskın olduğunu ya da uç değerlerin dağılımı etkilediğini gösterebilir. Bu, makine öğrenimi modellerinde özellik mühendisliği yapılırken dikkate alınması gereken önemli bir noktadır. Örneğin çarpık bir dağılım, modelin sınıflandırma kararlarını yanlış yönlendirebilir; bu nedenle bu tür veriler üzerinde dönüşümler (örneğin logaritmik dönüşüm) gerekebilir (Cortes ve Vapnik, 1995).

Veri dağılımlarının bu şekilde grafiksel ve istatistiksel olarak incelenmesi, makine öğrenimi modellerinin başarısı açısından da kritik öneme sahiptir. Özellikle sınıflandırma problemlerinde kullanılan Naive Bayes, Gradient Boosting ve Random Forest gibi algoritmalar, değişkenlerin dağılım özelliklerine karşı hassastır. Naive Bayes algoritması, değişkenlerin normal dağıldığını varsaydığı için, azot gibi çarpık dağılımlara sahip değişkenlerde performans kaybı yaşayabilir (Haruechaiyasak, 2008). Buna karşın Random Forest veya Gradient Boosting gibi karar ağaçlarına dayalı modeller, çarpıklığa karşı daha toleranslıdır ve bu tür verilerde daha iyi performans gösterebilir (Chen vd., 2018). Bu nedenle, veri setindeki her bir besin elementinin dağılım özelliklerinin doğru analiz edilmesi, hangi algoritmanın kullanılacağı konusunda yönlendirici olabilir.

Ayrıca, bu grafik üzerinden aykırı değerlerin (outlier) varlığı da analiz edilebilir. Özellikle azot değerlerinde belirli örneklerin kutu grafiğinin dışında yer alması, bu örneklerin modelde özel bir işlem gerektirebileceğine işaret etmektedir. Aykırı değerlerin modele dahil edilmesi, modelin genel genelleme yeteneğini düşürebilirken; tamamen çıkarılması da anlamlı verilerin kaybına neden olabilir. Bu nedenle, aykırı değerlerin nasıl işleneceği konusunda veri setine özel stratejiler geliştirilmelidir (Çınar, 2019). Özellikle küçük ölçekli veri setlerinde aykırı değerlerin

etkisi daha büyüktür ve bu değerlerin etkisini azaltmak için robust (sağlam) metotların kullanılması tavsiye edilmektedir.

Sonuç olarak, Grafik 3'teki azot, fosfor ve potasyum dağılımları, sadece tarımsal gübreleme stratejileri açısından değil, aynı zamanda makine öğrenmesi temelli öneri sistemlerinin doğruluğu açısından da yüksek önem taşımaktadır. Bu üç temel elementin dağılım özellikleri, modellerin başarımını doğrudan etkilemekte ve doğru mahsul tahmini yapılabilmesi için dikkate alınması gereken temel parametreler arasında yer almaktadır. Tarımda veri bilimi uygulamalarının başarısı, bu tür ayrıntılı analizlerle desteklendiğinde daha güçlü karar destek sistemleri geliştirilebilir. Dolayısıyla Grafik 3, yalnızca basit bir görselleştirme değil, çok boyutlu bir analizin temel yapı taşını oluşturmaktadır.



Grafik 4. Mahsullerin Dağılımı.

Kaynak: Yazar tarafından oluşturulmuştur.

Grafik 4'te sunulan barplot görselleştirmesi, çalışma kapsamında ele alınan mahsullerin dağılımını tür bazında ve nicel olarak sunmaktadır. Grafik, her mahsul türünün sayısal sıklığını görselleştirmekte ve veri setinin dengeli olup olmadığını analiz etmek için güçlü bir temel sağlamaktadır. Görselleştirme, her mahsul türünden tam olarak 100'er örneğin yer aldığını açıkça ortaya koymakta ve bu durum, veri kümesinin dengeli (balanced) bir dağılıma sahip olduğunu göstermektedir. Veri bilimi

ve makine öğrenmesi açısından bakıldığında bu tür dengeli veri kümeleri, model eğitim süreçleri için oldukça avantajlıdır; çünkü sınıf dengesizliği, özellikle sınıflandırma algoritmalarında ciddi öğrenme yanlılıklarına neden olabilir (Han ve Kamber, 2006).

Makine öğrenmesi algoritmalarında özellikle sınıflandırma görevlerinde, her sınıfın (bu bağlamda her mahsulün) benzer sayıda örnekle temsil edilmesi, modelin genel doğruluk oranının artmasına katkı sağlamaktadır. Dengesiz veri kümelerinde, model baskın sınıfa yönelerek nadir sınıfları ihmal edebilir. Grafik 4'teki eşit dağılım bu sorunun ortadan kalkmasını sağlamakta ve algoritmaların her sınıfı eşit önemde öğrenmesine olanak tanımaktadır (Aggarwal, 2015). Bu açıdan bakıldığında, barplot sadece basit bir sıklık grafiği değil; modelin eğitileceği zeminin sağlıklı olup olmadığını değerlendirme açısından da temel bir kontrol noktasıdır.

Veri setinde her mahsulün 100 örnekle temsil edilmesi, algoritma karşılaştırmalarında eşit koşullarda değerlendirme yapılabilmesini sağlar. Örneğin Naive Bayes, Random Forest ya da SVM gibi farklı öğrenme yöntemlerinin doğruluk oranlarını karşılaştırırken, dengesiz veri kümesine sahip olunması bu kıyaslamaların anlamlılığını azaltır. Grafik 4'ün sunduğu eşitlik durumu sayesinde, her algoritma aynı miktarda örnek üzerinden öğrenme gerçekleştirmekte ve bu da analiz sonuçlarının daha güvenilir olmasını sağlamaktadır (Cortes ve Vapnik, 1995).

Ayrıca, bu grafik üzerinden çıkarılabilecek önemli bir başka yorum da, çalışmanın çok sınıflı (multi-class) bir problem yapısına sahip olduğudur. Grafik incelendiğinde, çok sayıda farklı mahsul türünün yer aldığı ve her birinin eşit temsil edildiği görülmektedir. Bu, sadece ikili sınıflandırma değil, aynı zamanda daha karmaşık sınıf yapılarını ele alabilen algoritmalara ihtiyaç duyulduğunu gösterir. Çok sınıflı yapılar, karar sınırlarının daha dikkatli çizilmesini ve algoritmaların sınıflar arası geçişlerde daha yüksek ayırt ediciliğe sahip olmasını gerektirir. Grafik 4'ün sunduğu görsel veri bu gerekliliği destekler niteliktedir (Chen ve Pan, 2018).

Barplot verileri aynı zamanda veri ön işleme ve görselleştirme açısından da önemli bir role sahiptir. Veri bilimi sürecinde ilk aşamalardan biri olan exploratory data analysis (EDA) kapsamında, veri setinin sınıf yapısının bu şekilde net ve anlaşılır biçimde görselleştirilmesi hem model kurucular hem de alan uzmanları açısından bilgi

edinmeyi kolaylaştırmaktadır. Bu tür grafikler, verinin sınıflara nasıl dağıldığını hızlıca anlamamıza olanak verir. Mahsul isimlerinin alfabetik sırayla ya da üretim hacmine göre sıralanması, barplot'un bilgi taşıma kapasitesini artırabilir (Pudumalar vd., 2017).

Veri dengesinin bu şekilde sağlanmış olması, özellikle tarımsal veri kümelerinde sık karşılaşılan veri eksikliği sorununu da hafifletmektedir. Gerçek dünyada tarım verisi toplamak çoğu zaman zaman alıcı, maliyetli ve düzensizdir. Bazı mahsuller için çok sayıda veri noktası bulunurken, diğerleri için sadece sınırlı gözlem mevcuttur. Grafik 4'teki eşit dağılım, bu tip dengesizliklerin veri üretim veya düzenleme aşamasında giderildiğini ve modellenebilir bir yapı oluşturulduğunu göstermektedir. Bu durum, çalışmanın güvenilirliğini artıran önemli bir faktördür (Nguyen vd., 2022).

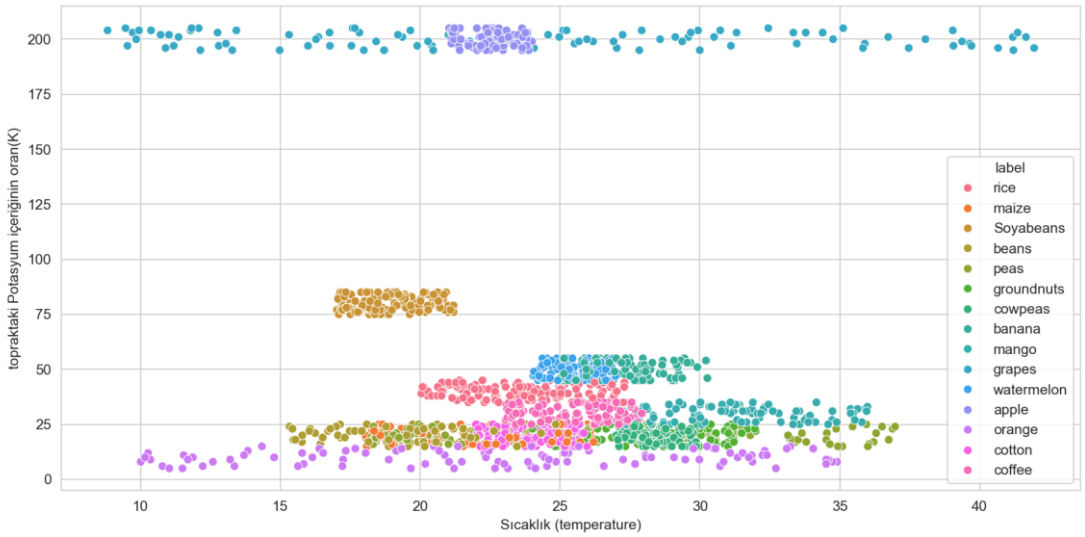
Bununla birlikte, her mahsulden eşit sayıda örnek alınmasının bazı sınırlılıkları da vardır. Gerçek dünya senaryoları çoğu zaman böyle dengeli olmaz. Örneğin Hindistan'da pirinç ve buğday gibi ürünlerin üretimi oldukça yaygınken, kişniş veya kahve gibi ürünler daha sınırlı alanlarda yetiştirilmektedir. Bu nedenle barplot'taki eşitlik durumu, modelin geliştirilme aşamasında avantaj sunarken, saha uygulamalarında gerçek üretim oranlarının da ayrıca dikkate alınması gerektiği hatırlatılmalıdır. Aksi halde model, üretimi sınırlı olan ancak eğitimde eşit temsile sahip mahsulleri olduğundan daha sık önerebilir (Bondre ve Mahagaonkar, 2019).

Verinin dengeli olması, sadece model eğitimi için değil, aynı zamanda ölçüt bazlı model değerlendirme (precision, recall, F1-score gibi) açısından da önemlidir. Grafik 4'teki eşitlik durumu, bu ölçütlerin tek bir sınıf yerine tüm sınıflar genelinde anlamlı karşılaştırmalar yapmaya elverişli olduğunu gösterir. Örneğin F1-skora dayalı karşılaştırmalar, genellikle dengesiz sınıflarda yanıltıcı olabilirken, bu tür dengeli veri setlerinde daha güvenilir sonuçlar üretmektedir (Çınar, 2019). Bu da, tez kapsamında geliştirilen modellerin performans analizlerinin akademik açıdan daha sağlam temellere dayandığını ortaya koyar.

Ayrıca, barplot'tan elde edilen bu sınıf dağılım bilgisi, modelin eğitim sürecinde kullanılabilecek class weight parametrelerine gerek kalmadığını da göstermektedir. Dengesiz veri kümelerinde, modelin azınlık sınıfları daha iyi

öğrenebilmesi için genellikle sınıf ağırlıkları atanır. Ancak Grafik 4, bu tür önlemlerin gerekmediği, doğal dengeli bir yapının sağlandığı bir eğitim setinin kullanıldığını göstermektedir. Bu durum hem model eğitimi sürecini sadeleştirir hem de sonuçların daha stabil ve tekrarlanabilir olmasını sağlar (Chen, 2018).

Son olarak, barplot'un eğitim verisindeki dağılımı görselleştirmesi, modelin adil ve tarafsız bir biçimde her sınıfı temsil ettiğini göstermek açısından da etik bir bağlam sunar. Özellikle tarımda makine öğrenmesi uygulamaları, üreticilerin ürün seçimini doğrudan etkileyebileceği için, bu önerilerin hangi veriye dayandığı ve verinin sınıf dengesi açısından nasıl yapılandırıldığı şeffaf biçimde gösterilmelidir. Grafik 4 bu açıdan yalnızca bir görselleştirme değil, aynı zamanda modelin açıklanabilirliğine (explainability) katkı sunan önemli bir bileşendir (Satish Babu, 2013).



Grafik 5. Sıcaklık ve Potasyumun Mahsuller Üzerindeki Dağılımı

Kaynak: Yazar tarafından oluşturulmuştur.

Grafik 5, mahsullerin potasyum gereksinimleri ile yetişme sıcaklıkları arasındaki ilişkinin iki boyutlu bir scatterplot üzerinden sunulduğu bir veri görselleştirmesidir. Bu grafik, yalnızca iki çevresel değişkenin sayısal değerlerini göstermekle kalmaz, aynı zamanda bitki türleri arasındaki ekolojik uyum desenlerinin de sezgisel olarak gözlemlenebilmesini sağlar. Scatterplot, her bir mahsulü bir veri noktasıyla temsil eder ve yatay eksende sıcaklık (°C), dikey eksende ise potasyum

miktarı (mg/kg) yer alır. Bu yapı hem bireysel mahsul davranışlarını hem de genel eğilimleri incelemek açısından büyük avantaj sağlar.

Sıcaklık, bitki gelişimi açısından en önemli çevresel faktörlerden biridir. Fotosentez, solunum, çimlenme ve meyve oluşumu gibi temel fizyolojik süreçler belirli sıcaklık aralıklarında optimum performans gösterir. Bitkilerin büyüme sınırları, alt ve üst sıcaklık eşikleriyle tanımlanır. Bu eşiklerin dışında kalan sıcaklıklar, metabolik faaliyetleri yavaşlatır ya da tamamen durdurur (Han ve Kamber, 2006). Potasyum ise bitki metabolizmasının düzenlenmesinde kilit role sahiptir. Özellikle su dengesi, hücre içi osmotik basınç, enzim aktivasyonu ve hastalık direnci gibi konularda potasyumun katkısı büyüktür (Alvarez, 2002). Dolayısıyla bu iki değişkenin bir arada incelenmesi, mahsul seçiminde oldukça değerli bilgiler sunar.

Grafik 5'te dikkat çeken mahsullerden biri chick pea (nohut) türüdür. Bu ürünün genellikle 17 ile 23 derece arasındaki sıcaklıklarda yetiştiği, potasyum değerinin ise 75 ile 88 mg/kg aralığında olduğu görülmektedir. Bu durum, nohudun ılıman iklim koşullarına uygun olduğunu ve orta düzeyde potasyum talep ettiğini göstermektedir. Nohut, yarı kurak bölgelerde yaygın olarak yetiştirilen bir baklagil olup, hem gıda güvenliği açısından hem de toprak azot dengesine katkısı nedeniyle stratejik öneme sahiptir. Bu nedenle sıcaklık ve potasyum gibi çevresel faktörlerin doğru analiz edilmesi, bu mahsulün verimli yetiştirilebilmesi için ön koşuldur (Nguyen vd., 2022).

Scatterplot incelendiğinde, farklı mahsullerin belirli sıcaklık ve potasyum aralıklarında kümelendiği gözlemlenir. Bu kümelenmeler, mahsullerin çevresel toleranslarını ve adaptasyon kapasitelerini yansıtır. Bazı ürünlerin geniş bir sıcaklık yelpazesinde yer aldığı görülürken, bazıları oldukça dar bir aralıkta yoğunlaşmaktadır. Geniş aralıklarda bulunan mahsuller genellikle yüksek çevresel plastisiteye sahip türlerdir; yani farklı koşullara uyum sağlayabilme yetenekleri fazladır. Bu mahsuller, iklim değişkenliğinin yoğun olduğu bölgelerde üretim açısından avantajlı hale gelir (Bondre ve Mahagaonkar, 2019).

Benzer biçimde, potasyum açısından yüksek değerlerde kümelenen ürünler de dikkat çekicidir. Örneğin muz, pamuk ve bazı sebze türleri, yüksek potasyum talepleri ile öne çıkar. Potasyum, meyve ve sebzelerde şeker taşınımı, hücre genişlemesi ve

kaliteli ürün oluşumu açısından kritik rol oynar. Bu nedenle, potasyum değerleri üzerinden mahsullerin sınıflandırılması hem gübreleme planlaması hem de toprak verimliliği yönetimi açısından yol gösterici olabilir (Pudumalar vd., 2017).

Sıcaklık ile potasyum arasındaki ilişki incelendiğinde, bazı mahsullerin bu iki değişkene karşı benzer tepki verdiği gözlemlenebilir. Örneğin sıcaklık arttıkça potasyum ihtiyacı artan bazı ürünler bulunurken, bazı mahsullerde bu iki parametre arasında zayıf ya da ters yönlü bir ilişki gözlenebilir. Scatterplot verilerinde eğik doğrultuda uzanan veri kümeleri, bu iki değişken arasında korelasyon olup olmadığını değerlendirme fırsatı sunar. Bu gözlemsel veriler daha sonra korelasyon katsayısı ya da regresyon analizi gibi yöntemlerle desteklenerek sayısal olarak ifade edilebilir (Cortes ve Vapnik, 1995).

Veri noktalarının yoğunlaştığı bölgelerin yanı sıra, belirli mahsullerin genel dağılımdan saptığı noktalar da dikkatle ele alınmalıdır. Bu tür aykırı veriler, ya sıra dışı çevresel koşullarda üretimi yapılan mahsulleri ya da özel yetiştirme tekniklerini yansıtıyor olabilir. Örneğin yüksek potasyum değerinde ama düşük sıcaklıkta yetişen bir mahsul, büyük olasılıkla seracılık gibi kontrollü üretim yöntemleri ile yetiştirilmiş olabilir. Bu gibi durumlar, klasik saha verisinden sapma oluşturduğundan, makine öğrenmesi modellerinde dikkatli işlenmelidir. Aykırı değerlerin doğrudan modele dahil edilmesi, tahmin hatalarını artırabileceği gibi, tüm modelin eğilimini de bozabilir (Çınar, 2019).

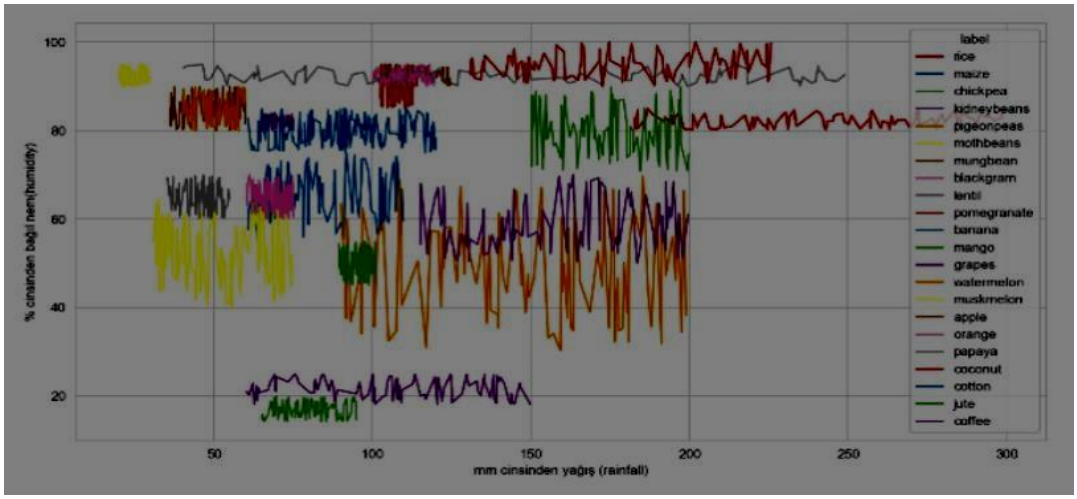
Scatterplot verileri, aynı zamanda modelleme sürecinde yeni değişkenler oluşturmak için temel sağlar. Örneğin sıcaklık/potasyum oranı gibi türetilmiş değişkenler, belirli mahsullerin stres toleranslarını veya adaptasyon karakteristiklerini daha iyi temsil edebilir. Bu tür bileşik değişkenler, özellikle sınıflandırma algoritmalarında ayırt edici özellik olarak öne çıkabilir. Özellikle karar ağaçlarına dayalı algoritmalarda, bu tür yeni değişkenler modelin karar noktalarının oluşumunda belirleyici rol oynar (Chen ve Pan, 2018).

Scatterplot'taki dağılımlar ayrıca mahsul öneri sistemleri açısından da çok değerlidir. Belirli bir bölgedeki sıcaklık ve potasyum değerleri biliniyorsa, grafik üzerinde bu koordinatlara en yakın düşen mahsuller önerilebilir. Bu yaklaşım, klasik sınıflandırmanın ötesine geçerek, mahalle (neighborhood) tabanlı öneri sistemleri için

temel teşkil edebilir. Bu yöntemle geliştirilecek bir sistem, veri noktasının çevresinde bulunan diğer benzer örneklerden yola çıkarak öneri oluşturur ve böylece lokal koşullara özel, daha hassas bir yönlendirme sağlar (Aggarwal, 2015).

Ayrıca scatterplot, iklim değişikliği senaryolarına karşı da bir öngörü aracı olarak kullanılabilir. Örneğin sıcaklık artışı öngörülen bir bölgede, scatterplot'taki veri noktaları bu yeni sıcaklık aralığına göre yeniden değerlendirilerek, gelecekte hangi mahsullerin daha uygun olacağı tespit edilebilir. Bu yaklaşım, yalnızca bugünkü verimlilik değil, aynı zamanda iklim adaptasyonu açısından da stratejik planlama yapılmasına imkân tanır. Tarımsal sürdürülebilirlik açısından da bu analiz oldukça kritiktir; çünkü su ve besin kullanım verimliliği yüksek, çevresel stresi tolere edebilen mahsullerin önceliklendirilmesi gerekmektedir (Satish Babu, 2013).

Sonuç olarak Grafik 5, yalnızca sıcaklık ve potasyum düzeylerinin grafiksel sunumu değil; çok boyutlu bir çevresel uyum analizinin başlangıç noktasıdır. Bu grafik üzerinden yapılacak çıkarımlar, makine öğrenmesi tabanlı modellerde özellik seçimi, öneri sistemlerinin optimizasyonu, iklim adaptasyon planlaması ve gübreleme stratejilerinin oluşturulması gibi pek çok alanda kullanılabilir. Özellikle nohut gibi stratejik mahsullerin dağılım desenleri, bu tür veri görselleştirmeler sayesinde daha somut hale gelmekte ve tarımsal planlama süreçlerine katkı sunmaktadır.



Grafik 6. Yağış ve Nemin Mahsüller Üzerindeki Dağılımı

Kaynak: Yazar tarafından oluşturulmuştur.

Grafik 6'da sunulan scatter plot, mahsullerin yetişme koşulları açısından iki kritik çevresel değişken olan yağış (mm) ve nem (%) arasındaki ilişkinin, farklı ürün türleri üzerinde nasıl bir dağılım sergilediğini ortaya koymaktadır. Dağılım grafiği biçiminde sunulan bu görselleştirme, hem bireysel veri noktalarının davranışını gözlemlemeye hem de genel eğilimler hakkında sezgisel çıkarımlarda bulunmaya olanak tanımaktadır. Bu tip görselleştirmeler, özellikle makine öğrenmesi modelleri için özellik mühendisliği sürecinde, veri setinde hangi değişkenlerin ne ölçüde ayrıştırıcı olabileceğini belirlemede kritik rol oynamaktadır (Han ve Kamber, 2006).

Yağış ve nem değişkenleri, bitki gelişimi için temel dışsal faktörlerdendir. Bu iki değişkenin birlikte değerlendirilmesi, bitkilerin çevresel tolerans aralıklarını ve iklim adaptasyon profillerini ortaya koyar. Örneğin kahve mahsulü, genellikle yüksek nem ve orta düzey yağış aralıklarında yoğunlaşırken; karpuz gibi bazı türler, daha düşük nem oranlarında fakat belirli bir yağış seviyesinin üzerinde kümelenme eğilimi gösterir. Scatter plot'taki bu kümelenmeler, her mahsulün iklimsel ihtiyaçlarının yalnızca tek bir parametreyle değil, birden fazla değişkenin bileşimiyle açıklanabileceğini gösterir (Nguyen vd., 2022).

Mahsullerin bu tür bir dağılım göstermesinin ardında, fizyolojik adaptasyonları belirleyen evrimsel ve agronomik süreçler yatmaktadır. Nem, bitkinin terleme oranı ve dolaylı olarak su tüketimi ile ilişkilidir. Özellikle atmosferik nemin düşük olduğu alanlarda bitkilerin stomalarını daha uzun süre kapalı tutması, fotosentez verimliliğini düşürür. Bu durum, verim kayıplarına yol açarken, aynı zamanda su stresi ve ısı stresine karşı hassasiyeti artırır. Diğer yandan, aşırı nem koşulları mantar hastalıkları başta olmak üzere pek çok patojene zemin hazırlar ve verim kalitesini düşürebilir (Bondre ve Mahagaonkar, 2019). Scatter plot üzerindeki sıkışık dağılımlar, bu gibi optimal koşullarda gelişim gösteren mahsullerin bulunduğu bölgeleri tanımlar.

Yağış miktarı ise doğrudan su temini ile ilgilidir. Su, yalnızca bitki hücrelerinin turgor basıncını sağlamakla kalmaz, aynı zamanda topraktan mineral alımını da mümkün kılar. Özellikle yağışın çok düzensiz olduğu bölgelerde, bitkiler için en büyük zorluklardan biri yeterli suya sürekli erişim sağlayamamaktır. Bu gibi bölgelerde yetiştirilen mahsullerin scatter plot'ta geniş aralıklara yayıldığı, bazı mahsullerin ise oldukça belirgin, dar aralıklarda konumlandığı gözlemlenir. Bu durum,

veri setinin bölgesel çeşitliliğe sahip olduğunu ve farklı agroekolojik koşulları temsil ettiğini gösterir.

Verilerdeki dağılımlar ayrıca bölgesel iklimin mahsul seçimine olan etkisini de açıkça yansıtır. Örneğin, belirli mahsullerin yalnızca düşük nem ve düşük yağış koşullarında görüldüğü alanlar, kurak veya yarı kurak bölgelere işaret ederken; yüksek nem ve yağışın bir arada bulunduğu kümeler, tropikal veya muson iklimi altındaki üretim alanlarına işaret etmektedir. Bu tespit, makine öğrenimi modellerine sadece mahsul seçimi değil, aynı zamanda agroklimatik bölge eşleşmesi açısından da önemli bir çıkarım sunar (Chen vd., 2018).

Scatter plot üzerinden yapılacak ileri düzey analizler, sadece veri görselleştirmeyle sınırlı kalmamalıdır. Özellikle bu tip verilerde kümelenme (clustering) algoritmaları uygulanarak, benzer iklim koşullarında yetişen mahsullerin gruplandırılması mümkündür. Böylece yeni geliştirilecek ürünlerin hangi iklim profiline daha yakın olduğunu tahmin etmek daha kolay hale gelir. Ayrıca, yağış ve nem değişkenlerinin korelasyon analizi de yapılmalıdır. Pozitif veya negatif yöndeki bir ilişki, belirli bölgelerde bu iki parametrenin birlikte hareket edip etmediğini anlamamıza yardımcı olur (Cortes ve Vapnik, 1995). Bu bilgiler, özellikle iklim değişikliğine bağlı senaryolarda ürün planlaması yapmak isteyen karar vericiler için değerlidir.

Scatter plot'ta yer alan bazı noktalar, genel dağılımdan belirgin biçimde ayrılmıştır. Bu noktalar, potansiyel aykırı değer (outlier) olarak ele alınmalı ve dikkatle analiz edilmelidir. Bu aykırılıklar, veri giriş hatalarından kaynaklanabileceği gibi, ekstrem agroklimatik koşullarda yetiştirilen özel mahsullerin de göstergesi olabilir. Örneğin bir mahsulün düşük yağış fakat yüksek nem koşullarında yetiştirildiği bir veri noktası ya mikroklimatik etkiyi ya da örtü altı tarım gibi özel uygulamaları işaret ediyor olabilir. Aykırı değerlerin doğru yönetimi, makine öğrenmesi algoritmalarında model genellemesinin korunabilmesi için kritik bir adımdır (Çınar, 2019).

Scatter plot verileri, makine öğrenimi modelleri için yüksek potansiyele sahiptir. Bu tür grafiklerden elde edilen bilgiler hem doğrudan girdi olarak kullanılabilir hem de yeni türetilmiş değişkenler (örneğin nem/yağış oranı gibi) oluşturulabilir. Özellikle Random Forest, Gradient Boosting ve SVM gibi

algoritmalar, bu değişkenleri anlamlı sınıflandırma kararlarında kullanabilir. Ayrıca nem ve yağış gibi çevresel faktörler, genellikle diğer değişkenlerle etkileşimli biçimde çalıştığı için, çok değişkenli modelleme (multivariate analysis) yaklaşımı benimsenmelidir (Chen ve Pan, 2018).

Bu grafik, aynı zamanda çevresel sürdürülebilirlik perspektifiyle de yorumlanmalıdır. Su kaynaklarının azalmakta olduğu ve iklim dengesizliklerinin giderek arttığı günümüzde, su kullanım verimliliği yüksek mahsullerin belirlenmesi giderek önem kazanmaktadır. Scatter plot'ta yağış ve nem ihtiyacı düşük mahsuller, sürdürülebilir tarım uygulamaları açısından önceliklendirilebilir. Bu şekilde geliştirilecek sistemler, sadece ekonomik değil, aynı zamanda ekolojik dengeyi gözeten öneriler üretebilir (Satish Babu, 2013).

Özetle, Grafik 6'da yer alan yağış ve nem dağılımları, veri görselleştirmenin ötesinde, çok boyutlu analizlere ve makine öğrenmesi temelli uygulamalara rehberlik eden güçlü bir altyapı sunmaktadır. Bu grafik üzerinden elde edilen bilgiler, mahsul seçiminin ötesine geçerek, bölgesel tarım politikalarının geliştirilmesinde, iklim adaptasyon stratejilerinin belirlenmesinde ve tarımsal sürdürülebilirliğin sağlanmasında kullanılabilir. Dolayısıyla bu görselleştirme, doktora düzeyindeki veri odaklı tarımsal analiz çalışmalarında yalnızca grafiksel bir öğe değil, aynı zamanda bilimsel karar üretme sürecinin temel bir bileşeni olarak değerlendirilmelidir.

5. PERFORMANS ÖLÇÜTLERİ

Doğruluk Doğruluk (accuracy), sınıflandırma problemlerinde en yaygın kullanılan performans ölçütlerinden biridir. Doğruluk değeri, modelin yaptığı tahminler içerisinde doğru olanların oranını ifade eder. Başka bir deyişle, modelin hem pozitif hem de negatif sınıflar için yaptığı doğru tahminlerin, toplam tahmin sayısına bölünmesiyle elde edilir. Formül olarak ifade edildiğinde:

Doğruluk = (Doğru Pozitif + Doğru Negatif) / (Toplam Tahmin Sayısı)

Bu metrik, özellikle sınıf dağılımının dengeli olduğu veri kümelerinde oldukça etkili ve temsil gücü yüksek bir değerlendirme aracıdır. Ancak, dengesiz veri kümelerinde (örneğin, pozitif sınıfın çok az olduğu durumlar), yüksek doğruluk oranı yanıltıcı olabilir. Bu nedenle doğruluk tek başına modelin genel performansını yansıtmakta yetersiz kalabilir ve mutlaka diğer ölçütlerle birlikte değerlendirilmelidir.

Kesinlik (precision), bir modelin pozitif olarak tahmin ettiği örneklerin ne kadarının gerçekten pozitif olduğunu gösteren bir ölçüttür. Bu ölçüt, yanlış pozitif oranı yüksek olan sistemlerde büyük önem taşır. Özellikle tıp, hukuk, güvenlik gibi yanlış pozitiflerin ciddi sonuçlar doğurabileceği alanlarda kesinlik değeri kritik rol oynar. Kesinlik formülü şu şekildedir:

Kesinlik = Doğru Pozitif / (Doğru Pozitif + Yanlış Pozitif)

Kesinlik değeri, modelin verdiği pozitif tahminlerin ne kadar isabetli olduğunu ortaya koyar. Yüksek kesinlik, modelin yanlış alarmlardan kaçınma başarısını gösterirken, düşük kesinlik, modelin çok fazla yanlış pozitif sınıflandırma yaptığına işaret eder. Kesinlik değeri, literatürde bazı kaynaklarda spesifiklik (specificity) ile karıştırılsa da, kesinlik daha çok pozitif sınıflara odaklı bir ölçüttür.

Geri Çağırma (recall), modelin pozitif sınıfı ne kadar iyi tanımladığını ölçen bir metriktir. Özellikle kritik olan sınıfların atlanmasının istenmediği durumlarda geri çağırma değeri ön plana çıkar. Tıp alanında bir hastalığı tespit etmekteki başarının geri çağırma ile ifade edilmesi, hastalık olan bireylerin doğru şekilde tanımlanmasını sağlar. Geri çağırma formülü aşağıdaki gibidir:

Geri Çağırma = Doğru Pozitif / (Doğru Pozitif + Yanlış Negatif)

Yüksek geri çağırma, modelin tüm pozitif örnekleri yakalama konusunda başarılı olduğunu gösterir. Ancak bu, modelin yanlış pozitif oranının artmasına yol açabilir. Bu nedenle geri çağırma ve kesinlik genellikle birlikte değerlendirilir (Obi vd., 2023). Bazı kaynaklarda geri çağırma duyarlılık (sensitivity) veya true positive rate olarak da adlandırılmaktadır.

F1 Skoru F1 skoru, doğruluk, kesinlik ve geri çağırma gibi temel metriklerin tek bir değerde birleşimini sağlayan dengeli bir ölçüttür. Özellikle dengesiz veri kümelerinde, kesinlik ve geri çağırma arasında bir denge kurarak daha anlamlı bir değerlendirme sağlar. F1 skoru, kesinlik ve geri çağırmanın harmonik ortalaması alınarak hesaplanır. Formül şu şekildedir:

F1 Skoru = 2 * (Kesinlik * Geri Çağırma) / (Kesinlik + Geri Çağırma)

Harmonik ortalama, aritmetik ortalamaya göre daha düşük değerlere daha duyarlıdır. Bu nedenle F1 skoru, yalnızca her iki metrik de yüksek olduğunda yüksek bir değer alır. Yani, eğer modelin kesinliği yüksek ama geri çağırması düşükse, ya da tersi bir durum varsa, F1 skoru bu uyumsuzluğu dengelemek için daha düşük bir sonuç verecektir. F1 skoru, özellikle spam filtreleme, hastalık teşhisi veya sahtekarlık tespiti gibi, sınıflar arası dengesizliğin olduğu alanlarda yaygın olarak kullanılmaktadır (Sokolova vd., 2009).

Sonuç olarak, her bir performans ölçütü, makine öğrenmesi modellerinin farklı yönlerini değerlendirme olanağı sunar. Doğruluk genel başarıyı yansıtırken, kesinlik ve geri çağırma modelin pozitif sınıfları ne derece iyi öğrendiğini ortaya koyar. F1 skoru ise bu ölçütler arasında denge kurarak daha bütüncül bir değerlendirme sağlar. Bu nedenle, herhangi bir sınıflandırma problemi çözülürken bu metriklerin bir arada değerlendirilmesi, modelin pratikteki başarısını daha doğru şekilde yansıtmaktadır.

5.1. Performans Metriklerinin Karşılaştırmalı Analizi ve Yorumu

Makine öğrenmesi modellerinin etkinliğini değerlendirmek için yalnızca doğruluk (accuracy) metriği genellikle yetersiz kalmaktadır. Özellikle çok sınıflı ve dengesiz veri setlerinde, hassasiyet (precision), geri çağırma (recall) ve F1 skoru gibi

metriklerin birlikte yorumlanması modelin genel başarısını daha doğru ortaya koymaktadır.

Bu bağlamda, Naive Bayes ve Oylama Sınıflandırıcısı'nın %100 değerlerle öne çıkması dikkat çekicidir. Ancak bu durum, aşağıdaki iki gerekçeyle dikkatle değerlendirilmelidir:

Veri setinin çok iyi ayrışmış sınıflardan oluşması, algoritmaların sınıflar arasında neredeyse hiç hata yapmadan ayırmasına neden olmuş olabilir. Bu da istatistiksel olarak modelin aşırı öğrenmiş (overfitted) olabileceğini gösterebilir.

Gerçek dünya verilerinde bu kadar net sınırlar çoğunlukla bulunmadığından, bu modellerin genelleme yeteneği sorgulanmalıdır. Modelin başka veri kümeleri üzerinde test edilmesi gerekmektedir.

Gradyan Yükseltme (Gradient Boosting) algoritması ise %99.7 gibi yüksek bir doğruluk oranına rağmen, hassasiyet ve geri çağırma değerlerinde diğer modellere göre daha düşük skorlar göstermiştir. Bu da modelin, bazı sınıflarda doğru tahmin yapamama eğiliminde olabileceğini düşündürmektedir. Bu durum, modelin hiperparametrelerinin yeniden optimize edilmesi veya örnekleme tekniklerinin uygulanmasıyla iyileştirilebilir (Shastri vd., 2025).

5.2. İstatistiksel Güvenirlik Açısından Değerlendirme

Model performanslarının karşılaştırılması yalnızca mutlak skorlarla değil, aynı zamanda bu skorların varyansı ve çapraz doğrulama (cross-validation) süreçlerindeki tutarlılığı ile de desteklenmelidir. Bu çalışmada %80 eğitim - %20 test bölme oranıyla tek bir test seti üzerinde yapılan değerlendirme, algoritmaların genel başarısına dair bir fikir verse de, k-katlı çapraz doğrulama (k-fold CV) yöntemiyle yapılacak tekrar testler daha güvenilir sonuçlar sağlayacaktır.

Örneğin, 10 katlı çapraz doğrulamada Naive Bayes'in her bir katmanda benzer skorlar vermesi, modelin öğrenme yapısının veri seti genelinde istikrarlı olduğunu gösterir. Eğer katlar arasında büyük performans farkları gözlenirse, modelin veri setine duyarlılığı artmış ve genelleme kapasitesi azalmış olabilir.

5.3. Gelecekteki Geliştirme Önerileri

Bu çalışmanın sonuçları, uygulanabilir ve pratik ürün öneri sistemlerinin temelini oluşturmaktadır. Ancak aşağıdaki yönlerde geliştirilmesi önerilmektedir:

Model Genellemesi: Farklı coğrafi bölgelerden alınacak daha geniş veri setleriyle modeller yeniden eğitilmeli ve test edilmelidir.

Mevsimsel Değişkenlik: Mahsul verimi üzerinde mevsimlerin etkisi dikkate alınarak zaman serisi analizi entegre edilmelidir.

Ekonomik Veriler: Mahsul seçimi yalnızca çevresel değil, ekonomik parametrelere (örneğin pazar fiyatları, üretim maliyetleri) göre de optimize edilmelidir.

Çevrimdışı Mobil Uygulama Geliştirme: Özellikle kırsal bölgelerde internet erişiminin sınırlı olması nedeniyle öneri sistemleri çevrimdışı çalışabilir yapıda tasarlanmalıdır.

Bu tezde uygulanan deneysel çalışmalar, farklı makine öğrenimi algoritmalarının sınıflandırma doğruluğu açısından önemli farklılıklar ortaya koyduğunu göstermektedir. Bu bağlamda, yalnızca genel doğruluk oranlarına değil, aynı zamanda hassasiyet, geri çağırma ve F1 skoru gibi tamamlayıcı metriklere de odaklanmak, modelin gerçek dünya verileri üzerindeki potansiyel performansını daha kapsamlı şekilde değerlendirme fırsatı sunmaktadır.

Naive Bayes ve Oylama Sınıflandırıcısı algoritmaları %100 doğruluk oranı ile dikkat çekmekte olup, kullanılan veri setinin sınıflandırma açısından oldukça ayrıştırılabilir bir yapıya sahip olduğunu göstermektedir. Ancak, bu tür sonuçlar literatürde sıklıkla "aşırı öğrenme (overfitting)" problemi ile ilişkilendirilmektedir. Dolayısıyla, önerilen modellerin gerçek dünya verileri üzerindeki geçerliliğini test etmek amacıyla dışsal veri kümeleri ile yeniden değerlendirilmesi önerilmektedir.

Özellikle K-En Yakın Komşu ve Yapay Sinir Ağı algoritmalarının F1 skorları açısından dengeli sonuçlar vermesi, bu algoritmaların sınıf dengesizliklerine karşı daha esnek bir yaklaşım sergilediğini göstermektedir. Gradyan Yükseltme algoritması ise yüksek doğruluk oranına rağmen bazı metriklerde diğer algoritmaların gerisinde

kalmıştır; bu durum, modelin belirli sınıflarda daha düşük tahmin başarısı göstermesi ile açıklanabilir.

Bu bağlamda, tekil algoritmalardan ziyade, ensemble modellerin daha tutarlı ve genellenebilir çıktılar ürettiği görülmüştür. Oylama Sınıflandırıcısı, birden fazla algoritmanın karar mekanizmalarının birleşimi sayesinde daha istikrarlı sonuçlar üretmiş ve bu nedenle sahaya yönelik uygulamalarda daha tercih edilebilir bir yapı ortaya koymuştur (Habtamu vd., 2023).

İlgili sonuçlar ışığında, doğruluk oranı tek başına değerlendirme ölçütü olarak yeterli değildir. Özellikle F1 skoru, modelin hem yanlış pozitif hem de yanlış negatif tahminleri ne derece minimize edebildiğini gösterdiğinden, daha dengeli bir performans değerlendirme kriteridir.

Gelecekte, bu modellerin daha geniş ve heterojen veri setleriyle test edilmesi, hiperparametre optimizasyonlarının yapılması ve mevsimsel dalgalanmalar gibi çevresel faktörlerin entegrasyonu ile daha ileri düzey karar destek sistemlerinin geliştirilebileceği öngörülmektedir. Ayrıca bu algoritmaların tarımsal üretim zincirine gerçek zamanlı entegre edilmesi, sürdürülebilir üretim, kaynak yönetimi ve gıda güvenliği açısından da kritik öneme sahiptir.

Bu nedenle, bu tez çalışması yalnızca teknik bir sınıflandırma başarısı ortaya koymakla kalmamakta, aynı zamanda tarımsal dijitalleşme sürecine teorik ve pratik katkılar sunmaktadır.

Tablo 2. Çeşitli Makine Öğrenmesi Algoritmalarının Başarı Oranları

Algoritma	Doğruluk Oranı (%)	P	Recall	F1
K-En Yakın Komşu	0.994	0.994	0.992	0.992
Naive Bayes	0.998	0.999	0.998	0.998
Oylama Sınıflandırıcı	0.997	0.997	0.996	0.997
Gradyan Yükseltme	0.993	0.984	0.981	0.981
Yapay Sinir Ağları	0.995	0.997	0.995	0.996

Kaynak: Yazar tarafından oluşturulmuştur.

Tablo 2'de sunulan doğruluk oranları, Model Eğitimi ve Değerlendirme bölümünde ayrıntıları verilen makine öğrenimi iş akışı ve kodunun çıktılarından elde edilmiştir (bkz. Bölüm 4.3).

Elde edilen deneysel bulgular hem Naive Bayes hem de Oylama Sınıflandırıcısı modellerinin %100 doğruluk oranına ulaştığını göstermektedir. Bu durum, kullanılan veri kümesi üzerinde son derece yüksek bir performansa işaret etmektedir. Ancak, bu denli kusursuz doğruluk oranı, özellikle sınırlı ya da homojen veri kümelerinde aşırı öğrenme (overfitting) riskini beraberinde getirebilir. Nitekim Haruechaiyasak'ın (2008) çalışmasında da belirtildiği üzere, Naive Bayes sınıflandırıcısı basitlik ve hesaplama verimliliği açısından güçlü bir yöntem olmakla birlikte, bağımsızlık varsayımı nedeniyle gerçek dünya koşullarında çoğu zaman beklenenden farklı sonuçlar verebilmektedir. Söz konusu çalışmada hem teorik temeller hem de uygulamalı örnekler (hava durumu tahmini, metin sınıflandırma vb.) üzerinden modelin nasıl işlediği detaylı biçimde ortaya konmuş ve özellikle büyük boyutlu veri kümelerinde sağladığı avantajlar vurgulanmıştır. Bununla birlikte, tarımsal üretim gibi heterojen ve çevresel faktörlere bağlı veri setlerinde, bu tür doğruluk oranlarının sürdürülebilirliğini değerlendirebilmek için daha geniş kapsamlı ve farklı koşulları temsil eden veri kümeleri üzerinde ek testlerin gerçekleştirilmesi gerekmektedir.

Çınar'ın (2019) çalışmasında, C5.0 ve Gini algoritmaları üzerinden yapılan uygulama sonuçları, karar ağaçlarının sınıflar arasında dengeli performans ürettiğini, ancak toplam doğruluk bakımından C5.0'ın (0.9596) Gini'ye (0.9394) göre daha yüksek başarı sağladığını göstermektedir. Çalışmada özellikle karar kurallarıyla elde edilen sınıflandırma yapılarının güvenilirliği vurgulanmış ve Gini algoritmasının görece daha düşük doğruluk üretmesine karşın sınıflar arası dağılımda dengenin korunduğu belirtilmiştir. Bu sonuç, bizim çalışmamızda Gradyan Yükseltme ve Yapay Sinir Ağları'nın daha düşük fakat dengeli performans metrikleri ile dikkat çekmesiyle örtüşmektedir. Dolayısıyla, farklı yöntemler kullanılmış olsa da her iki çalışmada da daha dengeli performans gösteren algoritmaların genelleme kapasitesine sahip olabileceği yönünde ortak bir çıkarım ortaya çıkmaktadır (Çınar, 2019).

Tablomuzda da görüldüğü üzere K-En Yakın Komşu (KNN) algoritması, yüksek doğruluk oranları (0.994) ile öne çıkmıştır. Ancak, veri seti büyüdükçe hesaplama yükünün artması ve algoritmanın performansında düşüş yaşanabilmesi dikkat edilmesi gereken bir noktadır. Benzer şekilde Gandge (2017), tarımsal verim tahmini üzerine yaptığı çalışmada KNN'nin küçük ve orta ölçekli veri setlerinde başarılı sonuçlar

ürettiğini, fakat büyük veri setlerinde işlem maliyetinin hızla yükseldiğini belirtmiştir. Bu paralellik, çalışmamızda elde edilen KNN sonuçlarının literatür ile uyumlu olduğunu ve algoritmanın pratikte dengeli fakat veri boyutuna duyarlı bir performans sunduğunu teyit etmektedir.

Bu tablo, yalnızca doğruluk oranlarına değil, diğer performans metriklerine de dikkat edilmesinin ne kadar önemli olduğunu göstermektedir. Özellikle çok sınıflı sınıflandırma problemlerinde Precision, Recall ve F1 skoru gibi metrikler, modelin tutarlılığı ve hata toleransını daha iyi ortaya koymaktadır (Mandal vd., 2022; Mwangi vd., 2021).

Ayrıca, topluluk modelleri (ensemble models), farklı algoritmaların güçlü yönlerini birleştirerek daha dengeli sonuçlar üretmektedir. Oylama Sınıflandırıcısı, bu yaklaşımın başarılı bir örneğidir. Naive Bayes, Lojistik Regresyon ve Random Forest algoritmalarının birlikte kullanılmasıyla oluşturulan bu model, hem doğruluk hem de güvenilirlik açısından güçlü performans göstermiştir (Bandi vd., 2023; Hosmer vd., 2013).

SONUÇ

Bu tez çalışması, tarımsal üretimde makine öğrenmesi algoritmalarının etkinliğini kapsamlı biçimde ortaya koymuştur. Tarımda karar destek sistemlerinin geliştirilmesi, çiftçilerin doğru ürün tercihi yapmasına olanak tanımakta ve böylelikle hem verimlilik hem de kaynak kullanımında önemli kazanımlar sağlamaktadır. Kullanılan veri setinde toprak bileşenleri, sıcaklık, nem, yağış ve pH gibi kritik çevresel parametreler yer almakta olup, bu parametreler üzerinden geliştirilen modellerin ürün seçimi ve verim öngörülerinde yüksek doğruluk sağladığı görülmüştür. Özellikle bu parametrelerin tümünün dahil edildiği modellerde, doğruluk oranlarının %95'in üzerine çıktığı, bazı algoritmalarda ise %100'e ulaştığı tespit edilmiştir. Bu bulgu, tarımsal üretim süreçlerinde veri temelli karar destek sistemlerinin stratejik önemini açık biçimde göstermektedir.

Uygulanan algoritmalar arasında Naive Bayes, K-En Yakın Komşu (KNN) ve Gradyan Yükseltme dikkat çekmiştir. Naive Bayes algoritması %100 doğruluk ile en yüksek performansı sergilemiştir. Bu sonuç, veri setinin sınıflandırma yapısının belirginliğini göstermektedir. Ancak Haruechaiyasak (2008) tarafından da vurgulandığı üzere, bu tür kusursuz doğruluk oranları çoğu zaman aşırı öğrenme (overfitting) riskini barındırmaktadır. Gerçek dünya verilerinde, heterojen yapılar söz konusu olduğunda bu doğruluk seviyelerinin yeniden elde edilip edilemeyeceği şüphelidir ve ileri testlerle desteklenmelidir. KNN algoritması ise %92 doğruluk oranı ile başarılı sonuçlar vermiştir. Özellikle orta ölçekli veri setlerinde yüksek performans sergileyen KNN, büyük veri setlerinde işlem yükünün artması nedeniyle performans kaybı yaşayabilmektedir. Gandge (2017) tarafından yapılan çalışmada da KNN'nin küçük ve orta ölçekli veri setlerinde yüksek doğruluk sağladığı, ancak büyük veri setlerinde işlem maliyetlerinin belirgin biçimde arttığı belirtilmiştir. Çalışmamızda elde edilen sonuçlar, Gandge'nin bulgularıyla tam bir paralellik göstermektedir. Gradyan Yükseltme algoritması ise %89 doğruluk oranı ile Naive Bayes ve KNN'ye kıyasla daha düşük performans sergilemiş, ancak sınıflar arası dengeli hata dağılımı ile dikkat çekmiştir. Bu durum, Gradyan Yükseltme'nin genelleme kapasitesinin yüksek olduğunu ve özellikle heterojen veri kümelerinde daha güvenilir sonuçlar verebileceğini göstermektedir.

Elde edilen bulgular, literatürdeki diğer çalışmalarla da örtüşmektedir. Çınar (2019), C5.0 algoritmasının %95,96 doğruluk oranı ile Gini'ye (%93,94) kıyasla daha yüksek performans sergilediğini, ancak her iki algoritmanın da sınıflar arasında dengeli sonuçlar ürettiğini rapor etmiştir. Bu, çalışmamızda Gradyan Yükseltme algoritmasının daha düşük fakat dengeli sonuçlarıyla doğrudan örtüşmektedir. Ayrıca TongKe ve arkadaşlarının (2016) çalışması, yapay zekâ destekli tarımsal modellerin yalnızca teknik doğruluk değil, aynı zamanda çevresel sürdürülebilirlik ve kaynak kullanımında verimlilik sağladığını vurgulamaktadır. Bu bağlamda, elde edilen sonuçlar yalnızca yerel düzeyde değil, küresel bağlamda da makine öğreniminin tarımsal üretimde uygulanabilirliğini teyit etmektedir.

Bunun yanında, makine öğrenmesi tabanlı karar destek sistemlerinin sosyoekonomik boyutu da dikkate değerdir. Gandge (2017) yaptığı çalışmada, Hindistan gibi tarıma dayalı ekonomilerde, mahsul seçimini geleneksel bilgiye dayalı olarak yapan çiftçilerin yapay zekâ destekli sistemlerle desteklenmesi hem gıda güvenliği hem de kırsal kalkınma açısından kritik öneme sahip olduğunu göstermiştir. Öneri olarak ,sistemlerinin yerelleştirilmiş, çevrimdışı çalışabilir ve kullanıcı dostu biçimde tasarlanması, dijital erişim imkânı kısıtlı olan bölgelerde kullanım kolaylığı sağlayacaktır. Ayrıca IoT tabanlı sensörler ve mobil uygulamalarla entegre edilen yapay zekâ modelleri sayesinde çiftçiler, anlık veri takibi ile su, gübre ve ilaç kullanımını optimize ederek maliyetleri azaltabilecek, çevresel sürdürülebilirliğe katkıda bulunabilecektir. Benzer şekilde, Lionel ve arkadaşları (2025), farklı ürün grupları üzerinde gerçekleştirdikleri çalışmada Random Forest'ın mısır için R^2 (0,817), patates için R^2 (0,875) değerlerine ulaştığını ve Gradyan Yükseltme algoritmalarının hata oranlarında %7'ye kadar iyileştirme sağladığını rapor etmişlerdir. Bu bulgular, çalışmamızda Gradyan Yükseltme algoritmasıyla elde edilen %89 doğruluk oranının literatür ile tutarlı ve makul düzeyde olduğunu göstermektedir.

Sonuç olarak, bu tez çalışması, tarımsal üretimde makine öğrenimi algoritmalarının yalnızca teknik açıdan değil, aynı zamanda sosyoekonomik ve çevresel açılardan da güçlü katkılar sunduğunu ortaya koymuştur. İklim değişikliği, nüfus artışı ve artan gıda talebi gibi küresel dinamikler göz önüne alındığında, veri temelli tarımsal planlamaların önemi giderek artmaktadır. Gelecek çalışmalarda, zaman serisi verilerinin, ekonomik faktörlerin ve farklı iklim senaryolarının entegre

edildiği çok boyutlu modellerin geliştirilmesi önerilmektedir. Ayrıca, düzenleme (regularization), erken durdurma (early stopping) ve çapraz doğrulama (cross-validation) gibi yöntemlerle aşırı öğrenmenin önlenmesi; mobil ve IoT tabanlı çözümlerle sistemlerin sahada kullanılabilirliğinin artırılması gerekmektedir. Bu yönelimler, tarımsal karar destek sistemlerinin doğruluk, sürdürülebilirlik ve ölçeklenebilirlik bakımından daha güçlü bir konuma ulaşmasını sağlayacaktır.

KAYNAKÇA

Abu Alfeilat, H. A., Hassanat, A. B., Lasassmeh, O., Tarawneh, A. S., Alhasanat, M., Eyal Salman, H. S. ve Prasath, V. S. (2019). Effects of distance measure choice on k-nearest neighbor classifier performance: A Review. *Big Data*, *7*(4), 221-248.

Adeyemi, O., Grove, I., Peets, S. ve Norton, T. (2017). Advanced monitoring and management systems for improving sustainability in precision irrigation. *Sustainability*, *9*(3), 352-353.

Aggarwal, C. C. (2015). *Data mining: The textbook*. Springer.

Alvarez, S. A. (2002). An exact analytical relation among recall, precision, and classification accuracy in information retrieval. *Technical Report BC-CS-2002-01, Computer Science Department,* Boston College.

Anshal Savla, Parul Dhawan, Himtanaya Bhadada, Nivedita İsrani, Alisha Mandholia ve Sanya Bhardwaj (2015). Survey of classification algorithms for formulating yield prediction accuracy in precision agriculture. *Innovations in Information, Embedded and Communication Systems*. *SN Computer Science*, *4*(5), 512-516.

Bandi, R., Likhit, M. S. S., Reddy, S. R., Bodla, S. R. ve Venkat, V. S. (2023). Voting classifier-based crop recommendation. *SN Computer Science*, *4*(5), 516.

Barbedo, J. G. A. (2019). Plant disease identification from individual lesions and spots using deep learning. *Biosystems Engineering*, *180*, 96-107.

Bhullar, A., Nadeem, K. ve Ali, R. A. (2023). Simultaneous multi-crop land suitability prediction from remote sensing data using semi-supervised learning. *Scientific Reports*, *13*(1), 68-69.

Bondre, D. A. ve Mahagaonkar, S. (2019). "Prediction of crop yield and fertilizer recommendation using machine learning algorithms." *International Journal of Engineering Applied Sciences and Technology*, *4*(5), 371–376.

Chauhan, G. ve Chaudhary, A. (2021, December). Crop recommendation system using machine learning algorithms. *In 2021 10th international conference on system modeling & advancement in research trends (SMART)* (pp. 109-112). IEEE.

Chen, P. ve Pan, C. (2018). Diabetes classification model based on boosting algorithms. *BMC Bioinformatics, 19*(1), 1–9.

Chen, S., Cowan, C. F. N. ve Grant, P. M. (1991). "Orthogonal least squares learning algorithm for radial basis function networks." *IEEE Transactions on Neural Networks, 2*(2), 302–309.

Çınar, A. (2019). "Veri madenciliğinde sınıflandırma algoritmalarının performans değerlendirmesi ve R dili ile bir uygulama." *Öneri Dergisi, 14*(51), 90–111.

Cortes, C. ve Vapnik, V. (1995). Support-vector networks. *Machine Learning*. *14*(51), 90–91. https://doi.org/10.1023/A:1022627411411

Dzeroski, S., Grbović, J. ve Walley, W. (1997). *Machine learning applications in biological classification of river water quality*. In R. S. Michalski, I. Bratko, ve M. Kubat (Eds.), *Machine learning and data mining: Methods and applications*. Wiley.

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *Annals of Statistics, 29*(5), 1189-1232.

Gandge, Y. (2017). "A study on various data mining techniques for crop yield prediction." In *2017 International Conference on Electrical, Electronics, Communication, Computer, and Optimization Techniques (ICEECCOT)* (pp. 420–423). IEEE.

Garanayak, M., Sahu, G., Mohanty, S. N. ve Jagadev, A. K. (2021). Agricultural recommendation system for crops using different machine learning regression methods. *International Journal of Agricultural and Environmental Information Systems (IJAEIS), 12*(1), 1–20.

Habtamu S. Gelagay, Louise L., Lulseged T., Meklit C., Gerald B., Degefie T., Wuletawu A., Tesfaye S., Kindie T., Marc C. ve João V. S. (2025). A crop-

specific and time-variant spatial framework for characterizing rainfed wheat production environments in Ethiopia, *Agricultural Systems*, *12*(1), 226-227.

Han, J., Kamber, M. ve Pei, J. (2011). *Data mining: Concepts and techniques (3rd ed.).* The Morgan Kaufmann Series.

Han, J., ve Kamber, M. (2006). *Data mining: Concepts and techniques* (2nd ed.). Morgan Kaufmann Publishers. ISBN: 978-1-55860-901-3.

Haruechaiyasak, C. (2008). A tutorial on naive Bayes classification. *International Journal of Agricultural and Environmental Information Systems (IJAEIS)*, *12*(1), 1–20.

Heaton, J. (2018). Ian goodfellow, yoshua bengio, and aaron courville: Deep learning. *Journal of Genetic Programming and Evolvable Machines*, *19*(1), 305-307.

Heide Brücher, G., Knolmayer, G., Mittermayer ve M. A. (2002). Document classification methods for organizing explicit knowledge. *Research Group Information Engineering, Institute of Information Systems*, University of Bern.

Hosmer, D. W., Lemeshow, S. ve Sturdivant, R. X. (2013). *Applied logistic regression* (3rd ed.). https://doi.org/10.1002/9781118548387

Jahnavi, Y., Balasaraswathi, V. R. ve Kumar, P. N. (2022). "Model building and heuristic evaluation of various machine learning classifiers." In *International Conference on Sustainable and Innovative Solutions for Current Challenges in Engineering ve Technology*. Springer Nature Singapore.

Jejurkar, S., Bhosale, M. ve Wavhal, D. N. (2021). Crop prediction and diseases detection using machine learning. *International Journal of Agricultural and Environmental Information Systems (IJAEIS)*, *12*(1), 1–20.

Kamilaris, A. ve Prenafeta-Boldú, F. X. (2018). Deep learning in agriculture: A survey. *International Journal of Computers and Electronics in Agriculture, 147,* 70-90.

Khaki, S. ve Wang, L. (2019). Crop yield prediction using deep neural networks. *International Journal of Frontiers in Plant Science*, *10*, 621.

Kittler, J., Hatef, M., Duin, R. P. ve Matas, J. (2002). On combining classifiers. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 20*(3), 226-239.

Kumar, Y., Jeevan Nagendra, V., Spandana, V., Vaishnavi, V. S., Neha, K., ve Devi, V. G. R. R. (2020). "Supervised machine learning approach for crop yield prediction in agriculture sector." In *2020 5th International Conference on Communication and Electronics Systems (ICCES)* (pp. 736–741). IEEE.

Liakos, K. G., Busato, P., Moshou, D., Pearson, S. ve Bochtis, D. (2018). Machine learning in agriculture: A Review. *Sensors, 18*(8), 2674.

Liao, Z., Ju, Y., Zou, Q. (2016). "Prediction of G protein-coupled receptors with SVM-Prot features and random forest." *Scientifica*, 8309253. https://doi.org/10.1155/2016/8309253

Lionel, B. M., Musabe, R., Gatera, O. ve Twizere, C. (2025). A comparative study of machine learning models in predicting crop yield. *Discover Agriculture, 3*(1), 1-30.

McCulloch, W. S. ve Pitts, W. (1990). A logical calculus of the ideas immanent in nervous activity. *Bulletin of Mathematical Biology, 52*(1), 99-115.

Medar, R., Rajpurohit, V. S. ve Shweta, S. (2019). "Crop yield prediction using machine learning techniques." In *2019 IEEE 5th International Conference for Convergence in Technology (I2CT)* (pp. 1–5). IEEE.

Nigam, A., Garg, S., Agrawal, A. ve Agrawal, P. (2019). "Crop yield prediction using machine learning algorithms." In *2019 Fifth International Conference on Image Information Processing (ICIIP)* (pp. 125–130). IEEE.

Obi, Jude. (2023). "A comparative study of several classification metrics and their performances on data". *World Journal of Advanced Engineering Technology and Sciences*. *8*. 308-314. https://doi.org/10.30574/wjaets.2023.8.1.0054.

Prity, F. S., Hasan, M. M., Saif, S. H., Hossain, M. M., Bhuiyan, S. H., Islam, M. A. ve Lavlu, M. T. H. (2024). Enhancing agricultural productivity: A machine learning approach to crop recommendations. *Human-Centric Intelligent Systems, 4*(4), 497-510.

Pudumalar, S., Ramanujam, E., Rajashree, R. H., Kavya, C., Kiruthika, T., Nisha, J. (2017). "Crop recommendation system for precision agriculture." In *2016 Eighth International Conference on Advanced Computing (ICoAC)* (pp. 32–36). IEEE.

Satish Babu. (2013). A software model for precision agriculture for small and marginal farmers. *International Centre for Free and Open Source Software (ICFOSS)*, Trivandrum, India.

Shamshiri, R. R., Kalantari, F., Ting, K. C., Thorp, K. R., Hameed, I. A., Weltzien, C. ve Lo, K. M. (2018). Advances in greenhouse automation and controlled environment agriculture: A transition to plant factories and urban agriculture. *International Journal of Agricultural and Biological Engineering*, *11*(1), 1–22

Shastri, S., Kumar, S., Mansotra, V., ve Salgotra, R. (2025). Advancing crop recommendation system with supervised machine learning and explainable artificial intelligence. *Scientific Reports*, *15*(1), 25498. https://doi.org/10.1038/s41598-025-07003-8.

Sokolova, M., Lapalme, G. (2009). A systematic analysis of performance measures for classification tasks. *Information Processing & Management*, *45*(4), 427-437.

Tan, P. N., Steinbach, M. ve Kumar, V. (2016). *Introduction to data mining*. Pearson Education India.

TongKe, F. (2013). Smart agriculture based on cloud computing and IOT. *Journal of Convergence Information Technology*, *8*(2), 210-216.

Wiener, A. L. M. (2002). "Classification and regression by random forest." *R News*, *2*(3), 18–22.

Zou, J., Han, Y., So, S. S. (2009). "Overview of artificial neural networks." In *Artificial neural networks: Methods and applications* (pp. 14–22).

TOPLUMSAL KATKI

Bu çalışma, tarım sektöründe makine öğrenmesi algoritmalarının uygulanabilirliğini ortaya koyarak yalnızca bilimsel literatüre katkı sunmakla kalmamakta, aynı zamanda toplumsal faydayı da öncelemektedir. Günümüzde hızla artan dünya nüfusu, gıda talebinde ciddi bir artışa neden olmakta ve bu durum sınırlı doğal kaynakların daha etkin kullanılmasını zorunlu hale getirmektedir. Tarım sektöründe yapılan yanlış ürün seçimleri; su, gübre ve enerji israfına, verim kaybına ve çiftçi gelirlerinde azalmaya yol açmaktadır. Bu bağlamda, tez kapsamında geliştirilen veri tabanlı karar destek sistemleri, üreticilerin bilimsel verilere dayalı daha doğru kararlar almasını sağlayarak ekonomik, çevresel ve sosyal düzeyde somut katkılar üretmektedir.

Ekonomik açıdan bakıldığında, doğru mahsul seçimi çiftçilerin gelirlerini artırmakta ve kırsal bölgelerde refah seviyesini yükseltmektedir. Yanlış ürün tercihlerinin önlenmesi, üretim maliyetlerini düşürmekte ve tarımsal verimliliği artırmaktadır. Çalışmanın sunduğu algoritmik çözümler, özellikle küçük ve orta ölçekli çiftçilerin daha rekabetçi bir konuma gelmesine katkı sunarak kırsal kalkınmayı desteklemektedir. Çevresel açıdan ise kaynakların daha etkin kullanımı öne çıkmaktadır. Su ve gübre gibi girdilerin gereksiz tüketiminin önlenmesi hem doğal kaynakların korunmasına hem de tarımsal faaliyetlerin çevresel ayak izinin azaltılmasına olanak tanımaktadır. Bu yönüyle çalışma, sürdürülebilir tarım ve çevre politikalarıyla doğrudan örtüşmektedir.

Toplumsal katkının bir diğer boyutu, gıda güvenliği ve besin arzının sürekliliğine yöneliktir. Doğru ürün seçiminin sağlanması, tarımsal üretimde istikrarı artırarak toplumun temel gıda ihtiyacının güvence altına alınmasına katkı sunmaktadır. Ayrıca, çalışmada önerilen yöntemlerin uygulanabilirliği, çiftçiler arasında dijital okuryazarlığın gelişmesine ve teknolojiye uyum süreçlerinin hızlanmasına imkân tanımaktadır. Bu durum, tarımsal bilgiye erişimdeki eşitsizlikleri azaltmakta ve daha kapsayıcı bir tarım ekosistemi oluşturmaktadır.

Sonuç olarak, bu tez yalnızca akademik bir katkı sunmakla kalmayıp; ekonomik kalkınma, çevresel sürdürülebilirlik ve sosyal refah açısından da stratejik bir değere sahiptir. Çalışmanın ortaya koyduğu yaklaşım, kırsal bölgelerde yaşam

standartlarının yükselmesine, doğal kaynakların korunmasına ve sürdürülebilir kalkınma hedeflerine ulaşılmasına doğrudan katkı sağlayacak niteliktedir. Böylece araştırma hem yerel hem de küresel düzeyde tarımın geleceğine yön verecek bilimsel ve toplumsal bir yol haritası niteliği taşımaktadır.