DOKUZ EYLÜL ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

MAKİNE ÖĞRENMESİ ALGORİTMALARI İLE RÜZGAR TÜRBİNLERİNİN GÜÇ TAHMİNİ

Burcu KARAALİ

Şubat, 2023
İZMİR

MAKİNE ÖĞRENMESİ ALGORİTMALARI İLE RÜZGAR TÜRBİNLERİNİN GÜÇ TAHMİNİ

Dokuz Eylül Üniversitesi Fen Bilimleri Enstitüsü

Yüksek Lisans Tezi

İstatistik Anabilim Dalı, Veri Bilimi Programı

Burcu KARAALİ

Şubat, 2023

İZMİR

YÜKSEK LİSANS TEZİ SINAV SONUÇ FORMU

BURCU KARAALİ, tarafından **PROF.DR. SELMA GÜRLER** yönetiminde hazırlanan **"MAKİNE ÖĞRENMESİ ALGORİTMALARI İLE RÜZGAR TÜRBİNLERİNİN GÜÇ TAHMİNİ"** başlıklı tez tarafımızdan okunmuş, kapsamı ve niteliği açısından bir Yüksek Lisans tezi olarak kabul edilmiştir.

……

Prof. Dr. Selma GÜRLER

Yönetici

……	……
Prof. Dr. Gözde ULUTAGAY	Dr. Öğr. Üye. Engin YILDIZTEPE
(Jüri Üyesi)	(Jüri Üyesi)

Prof. Dr. Okan FISTIKOĞLU

Müdür

Fen Bilimleri Enstitüsü

TEŞEKKÜR

Bu çalışma boyunca, bana yön gösteren, sabırla motivasyon sağlayan ve kıymetli bilgisiyle desteklerini benden esirgemeyen sayın danışmanım Prof. Dr. Selma Gürler'e, aynı zamanda son zamanların en zorlu yıllarında bana olan inancı ve sonsuz desteği ile çalışma boyunca yanımda olan sevgili eşim İdris Karaali'ye ve beni yüreklendiren aileme çok teşekkür ederim. Çalışmanın uygulama sürecinde, ENERCON Servis LTD. ŞTİ. firmasına bizlere sağladığı rüzgar türbini işleyişi hakkındaki bilgiler ve rüzgar gücünü tahmin ederken algoritmaların performansını ölçmek için kullanılan verileri analiz etmemizde sağladıkları kaynaklar için teşekkür ederim.

Burcu KARAALİ

MAKİNE ÖĞRENMESİ ALGORİTMALARI İLE RÜZGAR TÜRBİNLERİNİN GÜÇ TAHMİNİ

ÖZ

Son yıllarda büyük bir ilgiyle teknik ve endüstriyel gelişmelerin bir sonucu olarak dünya çapında elektrik talebi her geçen gün artmaktadır. Ülkeler elektrik ihtiyaçlarını karşılamak için yenilenebilir enerji kaynaklarını verimli kullanmanın yollarını aramaktadır ve rüzgar enerjisi önemli ve yaygın bir yenilenebilir enerji kaynağıdır. Ayrıca son yıllarda makine öğrenmesi ve uygulamaları da ilgi görmektedir. Bu çalışma, gerçek bir rüzgar santralinden elde edilen geçmiş verileri kullanarak rüzgar gücü tahmini için bazı makine öğrenmesi yöntemlerini uygulamayı ve yenilenebilir enerji sektöründe stratejik kararlar vermeyi amaçlamaktadır. İncelenen bu yöntemler arasında k-en yakın komşuluk regresyonu, destek vektör regresyonu, rassal ormanlar regresyonu, gradyan arttırma makineleri ve aşırı gradyan arttırma makineleri bulunmaktadır. Bu çalışma ile, farklı makine öğrenimi algoritmalarının rüzgar türbinlerinin gerçek dünya verileri üzerindeki başarısı araştırılmış ve rüzgar enerjisi üretim tahmini için nihai modelin seçimi üzerine bu algoritmalar kıyaslanmıştır.

Anahtar Kelimeler: Makine öğrenmesi algoritmaları, rüzgar türbini, rüzgar gücünün tahmini

POWER PREDICTION OF WIND TURBINES WITH MACHINE LEARNING ALGORITHMS

ABSTRACT

The demand for electricity has been rising daily across the globe as a result of the enormous technical and industrial advancements in recent years. Countries have been looking for ways to efficiently use their renewable energy sources to meet their electrical needs, and wind energy is a significant and common renewable energy source. However, machine learning and its applications have more attention in recent years. This study aims to evaluate some machine learning methods for wind power prediction using historical data obtained from a real wind farm and to make strategic decisions in the renewable energy sector. Models that are examined include k-nearest neighbour regression, support vector regression, random forest regression, gradient boosting machines and extreme gradient boosting machines. In this study, the success of different machine learning algorithms on real world data of wind turbines is investigated and these algorithms are compared on the selection of the final model for wind power generation prediction.

Keywords: Machine learning algorithms, wind turbine, wind power prediction

İÇİNDEKİLER

Sayfa

YÜKSEK LİSANS TEZİ SINAV SONUÇ FORMU ii

TEŞEKKÜR iii

ÖZ iv

ABSTRACT v

ŞEKİLLER LİSTESİ viii

TABLOLAR LİSTESİ x

BÖLÜM BİR – GİRİŞ 1

1.1 Rüzgar Enerjisi 3

1.2 Rüzgar Türbini Tasarımı 6

1.3 Motivasyon ve Tez Çalışmasının Bölümleri 9

BÖLÜM İKİ – VERİ MADENCİLİĞİ SÜREÇLERİ VE İLGİLİ LİTERATÜR 11

2.1 Veri Madenciliği Tanımı ve Süreci 11

2.1.1 Veri Madenciliği Süreci 12

2.2 Makine Öğrenmesinde Kullanılan Yöntemler 13

2.2.1 Denetimli (Gözetimli) Öğrenme 13

2.2.2 Denetimsiz (Gözetimsiz) Öğrenme 15

2.3 Literatür Taraması 16

BÖLÜM ÜÇ – KULLANILAN MAKİNE ÖĞRENMESİ YÖNTEMLERİ 22

3.1 K-en Yakın Komşuluk Regresyonu (KNN)..... 23

3.2 Destek Vektör Regresyonu (SVR)..... 25

3.3 Rassal Ormanlar Regresyonu (RF)..... 29

3.4 Gradyan Arttırma Makineleri (GBM)..... 34

3.4.1 Hiperparametreler..... 36

3.5 Aşırı Gradyan Arttırma Makineleri (XGBoost)..... 37

3.5.1 Sistem Optimizasyonları..... 38

BÖLÜM DÖRT – RÜZGAR GÜCÜNÜN TAHMİNLENMESİ..... 40

4.1 Veri Seti ve Analizi..... 40

4.2 Veri Ön İşleme Prosedürleri..... 47

4.3 Sonuçların Değerlendirilmesinde Kullanılan Hata Fonksiyonları..... 50

4.4 Rassal Ormanlar (RF) Regresyonu için Bulgular..... 50

4.5 K-en Yakın Komşuluk (K-NN) Regresyonu için Bulgular..... 53

4.6 Destek Vektör Regresyonu (SVR) için Bulgular..... 54

4.7 Gradyan Arttırma Makinesi (GBM) için Bulgular..... 57

4.8 Aşırı Gradyan Arttırma (XGBoost) irüçin Bulgular..... 60

BÖLÜM BEŞ – SONUÇ..... 62

KAYNAKLAR..... 64

ŞEKİLLER LİSTESİ

Sayfa

Şekil 1.1 Türkiye'de yıllara göre kümülatif kurulu güç (TÜREB, 2023) 2
Şekil 1.2 Dünyada 2021 yılı deniz ve karada rüzgar enerjisinin kurulu gücü (GlobalWind Energy Council [GWEC], 2022)..... 2
Şekil 1.3 Türkiye'de rüzgar enerjisi potansiyel haritası (Özen, 2021) 5
Şekil 1.4 İşletmedeki rüzgar enerji santrallerinin bölgelere göre dağılımı (Özen, 2021) 5
Şekil 1.5 Rüzgar türbini tasarımı (ENERCON, 2021)..... 7
Şekil 1.6 ENERCON E-82 E3 hesaplanmış güç eğrisi (ENERCON, 2021) 8
Şekil 1.7 Rüzgar türbini parçalarının ayrıntılı görünüm (ENERCON,2021)..... 8
Şekil 1.8 Gerçekleşen gerçek zamanlı enerji üretimi (TEİAŞ, 2022)..... 9
Şekil 3.1 Makine öğrenmesi algoritmalarının sınıflandırma diyagramı 22
Şekil 3.2 Öklid uzaklık hesabı ve gösterimi 24
Şekil 3.3 Hiper düzlem ve destek vektörleri (Agarwal, 2021)..... 26
Şekil 3.4 Destek vektör regresyon grafiği (Buturache ve Stancu, 2021)..... 27
Şekil 3.5 Rassal ormanlar regresyonu akış diyagramı (Rodriguez-Galiano vd., 2016) 32
Şekil 3.6 Gradyan arttırma (GBM) algoritma adımları (Hastie vd., 2001)..... 36
Şekil 3.7 Karar ağacı algoritmalarının XGBoost'a evrimi..... 38
Şekil 4.1 Aylara göre gücün kutu grafiği 42
Şekil 4.2 Önerilen model için fonksiyonel diyagram..... 43
Şekil 4.3 Veri analizi ve tahmin yaparken gerçekleşen adımlar 43
Şekil 4.4 Rüzgar hızı ve güç arasındaki ilişki-güç eğrisi 44
Şekil 4.5 Aylara göre ortalama gerçekleşen güç (kW) 45
Şekil 4.6 SCADA 2019 ve 2020 verisi ile aylara göre ortalama rüzgar hızları (m/s) 45
Şekil 4.7 Türbinin hakim rüzgar yönü 46
Şekil 4.8 Güç (kW), rüzgar hızı ve rüzgar yönü arasındaki üç boyutlu (3B) görsel.. 46
Şekil 4.9 Rüzgar hızı ile ağırlıklandırılmış rüzgar yönü frekansları..... 47
Şekil 4.10 Rüzgar hızının aykırı gözlem tespiti 48
Şekil 4.11 Girdi ve çıktı parametreleri arasında ilişki matrisi 49

Şekil 4.12 Serpme diyagramı (scatter plot) ile girdi ve çıktı parametreleri arasındaki ilişki 49
Şekil 4.13 Rassal ormanlar regresyonu kullanılırken rüzgar hızı (m/s) ile türbinin güç üretimi (kW) arasındaki ilişkiyi gösteren dağılım grafiği 52
Şekil 4.14 Singh vd.(2021) rassal ormanlar regresyonu için çalışmasında elde ettiği bulgular 52
Şekil 4.15 Optimum k sayısının belirlenmesi 53
Şekil 4.16 K-NN regresyonu kullanılırken rüzgar hızı (m/s) ile türbinin güç üretimi (kW) arasındaki ilişkiyi gösteren dağılım grafiği 54
Şekil 4.17 Singh vd.(2021) k-en yakın komşuluk regresyonu için çalışmasında elde ettiği bulgular 54
Şekil 4.18 Optimum cost (maliyetin) belirlenmesi 56
Şekil 4.19 Destek vektör regresyonu kullanılırken rüzgar hızı (m/s) ile türbinin güç üretimi (kW) arasındaki ilişkiyi gösteren dağılım grafiği 57
Şekil 4.20 Ağaç sayısının belirlenmesi 58
Şekil 4.21 Gradyan arttırma makineleri kullanılırken rüzgar hızı (m/s) ile türbinin güç üretimi (kW) arasındaki ilişkiyi gösteren dağılım grafiği 59
Şekil 4.22 Singh vd.(2021) gradyan arttırma makinesi çalışmasında elde ettiği bulgular 59
Şekil 4.23 XGBoost kullanılırken rüzgar hızı (m/s) ile türbinin güç üretimi (kW) arasındaki ilişkiyi gösteren dağılım grafiği 61
Şekil 5.1 XGBoost doğrulama testi 63

TABLOLAR LİSTESİ

Sayfa

Tablo 1.1 Rüzgar türbinin teknik özellikleri 7
Tablo 2.1 Son beş yıl içinde geliştirilen bazı çalışmalar 21
Tablo 3.1 RF uygulama alanları (Fawagreh, Gaber ve Elyan, 2014) 33
Tablo 4.1 Veri setinin tanımlayıcı istatistikleri 41
Tablo 4.2 Model ayarlama işleminden sonraki çıktılar 51
Tablo 4.3 k=2 iken ve model ayarlama işleminden sonraki çıktılar 53
Tablo 4.4 Model svm ve model ayarlama işleminden sonraki çıktılar 56
Tablo 4.5 Model GBM ve model ayarlama işleminden sonraki çıktılar 58
Tablo 4.6 Model XGBoost ve model ayarlama işleminden sonraki çıktılar 60
Tablo 5.1 Modellerin tahmin performansının karşılaştırılması 62
Tablo 5.2 Doğrulama testi sonrası hata çıktıları 63

BÖLÜM 1
GİRİŞ

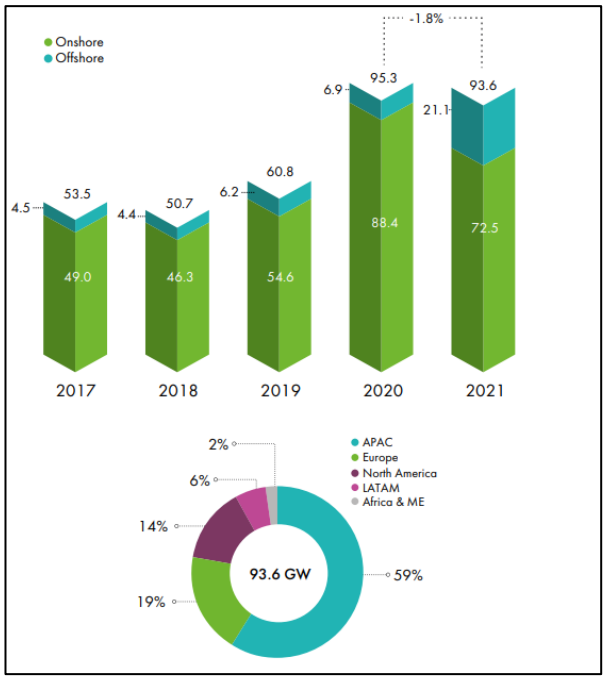
Günümüzde enerji, dünya ve insanların en önemli ihtiyaçlarından biri haline gelmiştir. Artan ihtiyaçlar ve bunlara yönelik geliştirilen üretim süreçleri nedeniyle enerji ihtiyacı ve enerji tüketimi de artmaktadır. Teknolojinin gelişmesi, sanayileşme ve yaşam standartlarının iyileşmesi ile birlikte her ülkenin enerji kaynakları, bu enerji kaynaklarının nasıl kullanıldığı, kullanım verimliliği, değerlendirilmesi, yönetimi ve geliştirilmesi dünya siyasi sürecini de şekillendirmektedir. Her ülke için enerji bağımsızlığı siyasi kararlılık katmaktadır. Tüm bu nedenlerle, gelecekte üretim planlamasında yenilenebilir enerji kaynaklarının kullanımının artırılmasına yönelik araştırmalar giderek önem kazanmaktadır (Karabağ vd., 2021).

Türkiye'nin rüzgar enerjisi potansiyelinin yüksek olduğu göz önüne alındığında, yenilenebilir enerji kaynaklarına talep artmaktadır. Rüzgar enerjisinin Türkiye'nin gelecekteki enerji üretimine vereceği katkıların önemli olacağı öngörülmektedir. Türkiye Rüzgar Enerjisi Birliği'ne (TÜREB) göre (Şekil 1.1), rüzgar enerjisinde Türkiye'nin kurulu gücü 2021 yılı sonunda toplam 11.101 MW'a ulaşmıştır (Türkiye Rüzgar Enerjisi Birliği (TÜREB), 2023). Türkiye'nin yedi coğrafi bölgesi genel olarak değerlendirildiğinde rüzgar türbinleri için en uygun bölgelerin Marmara, Ege ve Güneydoğu Anadolu bölgeleri olduğu, en fazla rüzgar potansiyeline sahip bölgelerin ise Çanakkale, İzmir, Balıkesir, Hatay ve İstanbul olduğu sonucuna ulaşılmıştır (İlkiliç, 2012).

Yıl	Kümülatif Kurulum MWm	Yıllık Kurulum MWm
1998	8.70	8.70
2000	18.90	10.20
2003	20.10	1.20
2006	66.45	46.35
2007	239.45	173.00
2008	483.75	244.30
2009	988.55	504.80
2010	1,375.80	387.25
2011	1,865.70	489.90
2012	2,355.35	489.65
2013	3,019.15	663.80
2014	3,834.80	815.65
2015	4,730.55	895.75
2016	6,130.65	1,400.10
2017	6,838.65	708.00
2018	7,404.05	565.40
2019	8,079.15	675.10
2020	9,489.85	1,410.70
2021	11,101.82	1,611.97

Şekil 1.1 Türkiye'de yıllara göre kümülatif kurulu güç (TÜREB, 2023)

Şekil 1.2'de 2021 yılında dünya çapında 93,6 GW yeni rüzgar kapasitesi eklenmiştir. Dünyada rüzgar enerjisi kurulu güç bakımından en fazla oran Asya-pasifik (APAC) bölgesindedir.



Şekil 1.2 Dünyada 2021 yılı deniz ve karada rüzgar enerjisinin kurulu gücü (GlobalWind Energy Council [GWEC], 2022)

Rüzgar türbinlerinin hızlı gelişim sürecinde, rüzgar türbinlerinin performansı teknoloji ile sürekli iyileştirilmektedir. Rüzgar gücüyle ilgili belirsizliği azaltmanın bir yolu, tahmin araçlarının kullanılmasıdır. Bu tür araçların geliştirilmesi 15 yıldan fazla sürmektedir (Giebel vd., 2011). Rüzgar enerjisi aralıklı ve değişken bir şekilde oluştuğundan, farklı zaman ölçeklerinde rüzgar enerjisinin nerede, ne zaman ve ne kadar kullanıldığını belirlememiz gerekir. Ek olarak, bir rüzgar türbininden elde edilebilecek üretim, yerel ve bölgesel rüzgar eğilimlerine, mevsimsel farklılıklara, çevredeki bitki örtüsüne, rakıma ve yakınlardaki diğer türbinlere bağlı olarak değişecektir. Bu, daha etkili ve verimli bir sistem uygulamasıyla sonuçlanır. Bu nedenlerle rüzgar türbinlerinin ürettiği rüzgar enerjisinin tahminine ihtiyaç duyulmakta ve doğru tahmin için veri madenciliği teknikleri bu aşamada ön plana çıkmaktadır. Veri madenciliği, makine öğrenimi, örüntü tanıma, istatistik, veri tabanı yönetim sistemleri, akıllı sistemler ve veri görselleştirmenin kesişiminden oluşmuş olup bu bağlamda gelişmektedir.

Bu çalışmada, rüzgar santrali tarafından üretilen güç tahmini için farklı makine öğrenmesi yöntemleri kullanılmış ve modeller incelenmiştir. İlgili modeller, Türkiye'nin İç Anadolu bölgesindeki bir kara rüzgar santralinin merkezi denetim ve veri toplama (SCADA (Supervisory Control and Data Acquisition)) sistemi tarafından toplanan 2019 ve 2020 geçmiş verileri kullanılarak oluşturulmuştur. Ortaya çıkan modeller, çeşitli performans kriterleri kullanılarak karşılaştırılmıştır. Önerilen model, karar vericilerin maliyetleri hesaplamasına ve bakım faaliyetlerini planlamasına yardımcı olur. Aynı zamanda, rüzgar türbini endüstrisindeki operasyonları desteklemek için uygulanabilecek teknik bir çözümün geliştirilmesine fırsat tanımaktadır.

1.1 Rüzgar Enerjisi

Rüzgar, yeryüzünün eşit olmayan ısınması ve soğuması sonucu ortaya çıkan kuvvetlerin etkisi ile oluşan hava hareketleridir. Rüzgar enerjisi ise, rüzgarı oluşturan hava akımının sahip olduğu hareket enerjisidir (Işık ve Ünal, 2020). Bu enerji türü,

akışkanın kütlesi m ve hızı U'nun bir fonksiyonu olan kinetik enerji olarak bilinir. Kinetik enerji eşitliği,

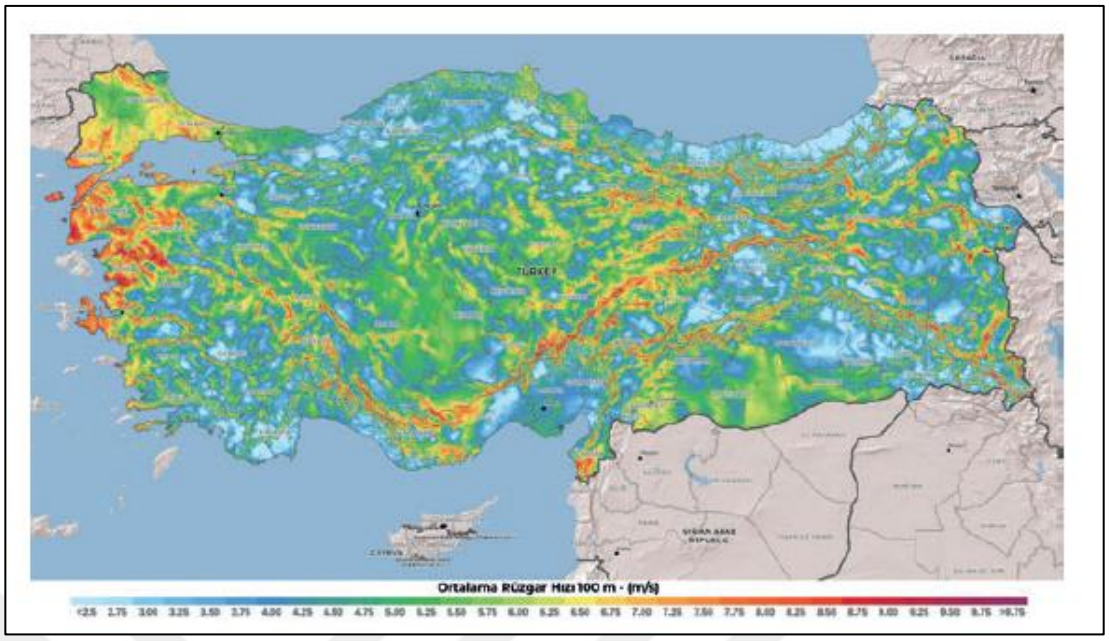
$$E = \frac{1}{2}mU^2 \tag{1.1}$$

olarak bilinir.

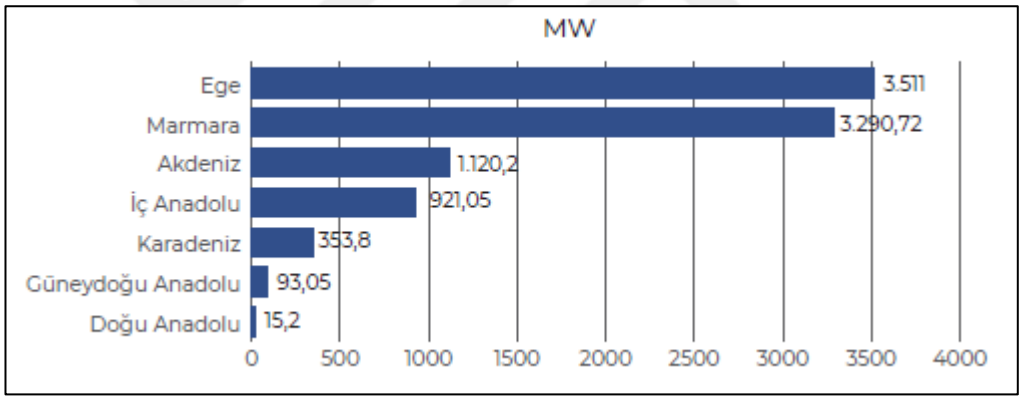
Rüzgardaki güç, kinetik enerji akış oranıdır. Rüzgar hızı U ve hava yoğunluğu ρ olan bir A alanını süpüren bir rüzgar türbini göz önüne alındığında, türbin tarafından üretilen maksimum rüzgar gücü,

$$P = \frac{1}{2}\rho A U^3 C_p \tag{1.2}$$

şeklinde hesaplanır. Türbin tarafından rüzgar enerjisi verimliliği güç katsayısı C_p ile ölçülür. Güç katsayısı, türbin tarafından çekilen gücün rüzgar kaynağının toplam gücüne oranını temsil etmektedir yani, $C_p = P_T/P_{wind}$. Bu nedenle, türbin gücü asla rüzgar gücü olan P_{wind}'den büyük olmamalıdır. Rüzgar gücü formülünde de görüldüğü gibi rüzgar hızı türbinlerin performansında büyük önem arz etmektedir. Şekil 1.3'te görüldüğü gibi Türkiye'de bölgelere göre ortalama rüzgar hızları görülmektedir. Bu ortalama rüzgar hızlarına bağlı olarak en çok kurulu güce sahip olan bölgeleri de Şekil 1.4'teki gibidir.



Şekil 1.3 Türkiye'de rüzgar enerjisi potansiyel haritası (Özen, 2021)



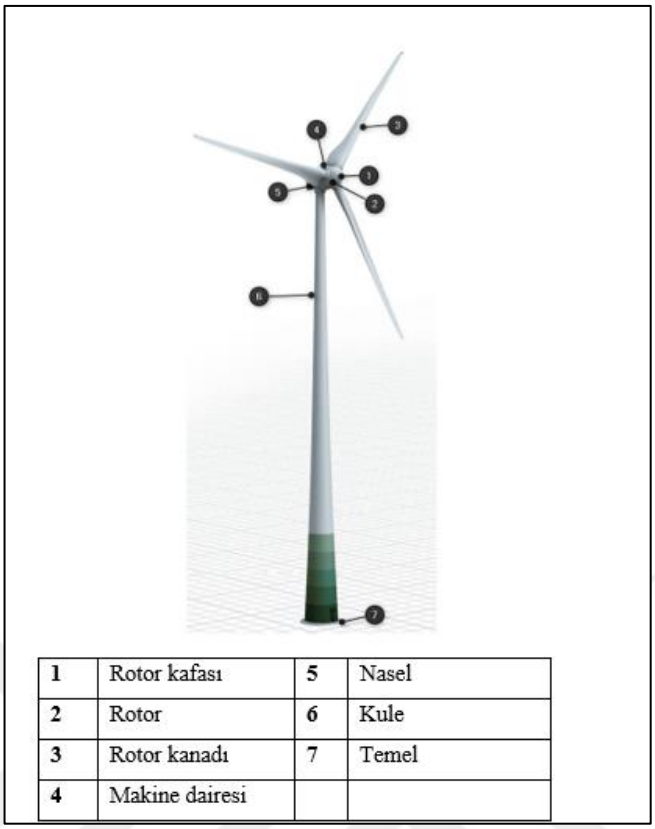
Şekil 1.4 İşletmedeki rüzgar enerji santrallerinin bölgelere göre dağılımı (Özen, 2021)

Rüzgar enerjisi potansiyel haritasına göre ülkemizde 5 MW/km² gücünde rüzgâr santralilerinin yerden 50 metre yükseklikte ve 7,5 m/s üzerinde rüzgâr hızlarına sahip alanlarda rüzgar santrali kurulabileceği kabul edilmiştir ve bu doğrultuda ülkemizin karadaki rüzgar potansiyelinin toplam 48 GW olduğu hesaplanmıştır. Yenilenebilir Enerji Kaynakları Destekleme Mekanizması (YEKDEM) ve Yenilenebilir Enerji Kaynak Sahası (YEKA) ihaleleri gibi girişimlerle son on yılda başta rüzgar enerjisi olmak üzere yenilenebilir enerjiye yapılan yatırımlar önemli ölçüde artmış olsa da ülkemizin karasal rüzgar enerjisi potansiyelini yeteri kadar kullanamadığı

görülmektedir. Ülkemiz karadaki rüzgar enerjisi potansiyelinin yaklaşık %15'ini son 10 yılda faaliyete geçen santraller ile değerlendirebilmiş olmasına rağmen kalan %85'lik potansiyel henüz değerlendirilmemiştir. Yerli ve milli enerji kaynaklarını kullanarak, enerjide dışa bağımlılığımızı azaltıp, gerekli elektriği temiz enerji kaynaklarından üretebilmek, yeni istihdam sektörlerinin ve iş alanlarının gelişmesi için önemli fırsatları beraberinde getirmektedir.

1.2 Rüzgar Türbini Tasarımı

Rüzgar türbinin yapısı Şekil 1.5'te göründüğü gibi ince, tepeye doğru daha da incelen bir kule üzerinde, kule ekseni etrafında dönebilen, nasel (nacelle) ve rotor kafasından oluşan, makine dairesinden oluşmaktadır. Rotor kafasında, üç rotor kanadı monte edilmiştir. Kanatlar uzunlamasına eksenleri etrafında döndürülerek ayarlanabilmektedir. Jeneratörün sabit duran bölümü (jeneratör statoru) makine dairesine aittir. Dönen bölümü (jeneratör rotoru) ise arada dişli kutusu olmadan doğrudan rotor kafasına bağlıdır. Makine dairesi, aerodinamik olarak şekillendirilmiş bir muhafazayla çevrilmiştir. Rüzgar türbini işletim sistemi ise, türbin, makine dairesini daima rüzgâr yönüne çevirmektedir ve rüzgar koşullarına daha iyi uyum sağlamak için değişken bir dönme hızıyla çalışmaktadır. Üretilen elektriğin gerilimi ve frekansı buna göre değişmektedir. Yerel elektrik iletim/dağıtım kurumunun elektrik şebekesine besleme yapmak için elektrik akımı, doğrultucular (makine dairesinde), invertörler ve transformatör (kule tabanında) üzerinden hazır duruma getirilmektedir. Bu çalışmada verisi kullanılan türbin üreticisi ENERCON E-82 E3 türbininin özellikleri Tablo 1.1'deki gibidir (ENERCON, 2021).



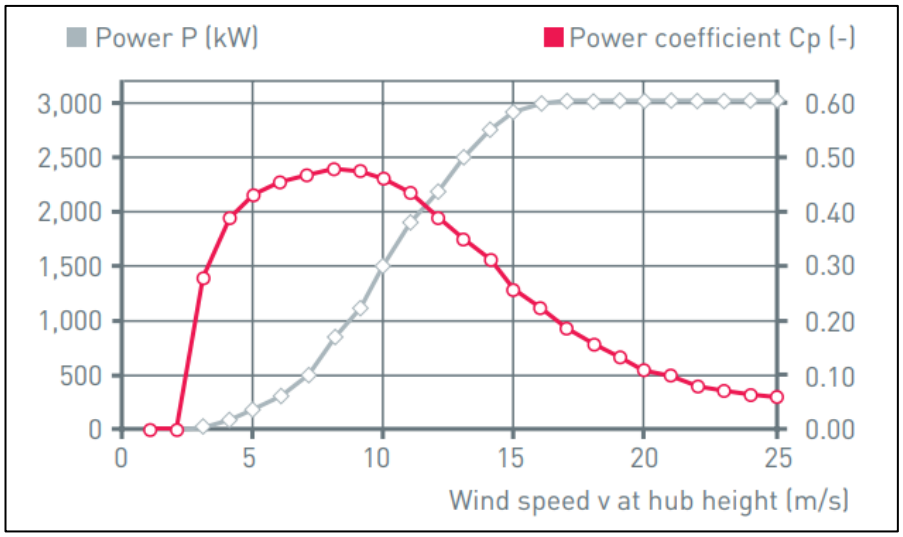
1	Rotor kafası	**5**	Nasel
2	Rotor	**6**	Kule
3	Rotor kanadı	**7**	Temel
4	Makine dairesi		

Şekil 1.5 Rüzgar türbini tasarımı (ENERCON, 2021)

Tablo 1.1 Rüzgar türbinin teknik özellikleri

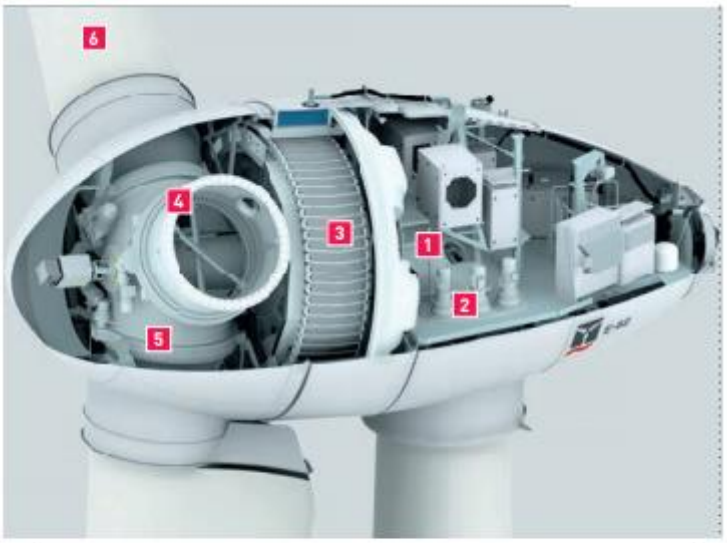
Özellik	**Değer**	**Birim**
Nominal güç	3000	kW
Nominal rüzgar hızı	16	m/s
Başlatma rügar hızı	2,5	m/s
Kapatma rüzgar hızı	25	m/s
Çap	82	m

Şekil 1.6'daki güç eğrisi grafiğinde görüldüğü üzere türbinin maksimum üreteceği güç 3000 kW olarak görülmektedir. Tablo 1.1'deki özelliklerden yola çıkarak nominal güce (3000kW) ulaştığı rüzgar hızı 16 m/s olarak görülmektedir. Aynı zamanda kapatma rüzgar hızı ise 25 m/s'dir. Buradan anlaşılacağı üzere, türbin maksimum kapasitesini 25 m/s rüzgar hızıyla sağlamaktadır ve 25 m/s üzeri rüzgar hızlarında ise nominal güçten daha fazla üretmemektedir.



Şekil 1.6 ENERCON E-82 E3 hesaplanmış güç eğrisi (ENERCON, 2021)

Şekil 1.7'de türbinin en üst bölmesindeki nasel (makine dairesi) kısmının içindeki parçaların ayrıntılı bir şekilde görünümü bulunmaktadır. Yapısal ve şekil olarak türbin çeşidine göre değişkenlik gösterse de genel olarak türbin işleyişi için kullanılan ana parçalardır.



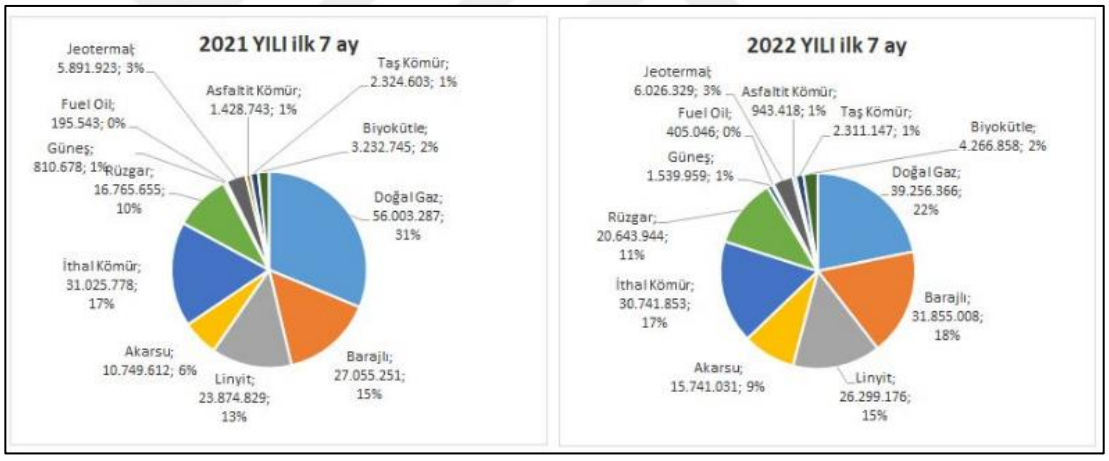
1	Ana taşıyıcı	**4**	Kanat adaptörü
2	Sapma sürüşü	**5**	Rotor göbeği
3	Dairesel jeneratör	**6**	Rotor kanadı

Şekil 1.7 Rüzgar türbini parçalarının ayrıntılı görünüm (ENERCON,2021)

1.3 Motivasyon ve Tez Çalışmasının Bölümleri

Son yıllarda, ucuz, bol enerji kaynakları arayışı önemli bir kamu politikası endişesi olmuştur. Enerji kıtlığı, sera gazları ve hava kirliliğinin getirdiği zorluklar, kamu ve özel sektörleri alternatif enerji kaynaklarının geliştirilmesine yatırım yapmaya zorlamıştır. Bununla birlikte alternatif enerji kaynaklarından son zamanlarda iklim ve küresel çapta krizin artmasıyla beraber yenilenebilir enerjiye bakış açıları eskiye oranla artmıştır. Bunu Şekil 1.8'deki veriler göstermektedir.

2022 yılı ilk 7 ayında, 2021 yılı aynı döneme göre %0,4 daha fazla elektrik üretimi gerçekleşmiştir. Doğalgaz ile elektrik üretimi %30 azalırken, rüzgar enerjisinde %23'lük, güneş enerjisinde %90'lık, biyokütlede ise %32'lik bir artış vardır. En büyük pay %22 ile doğalgaz ve %18 ile barajlı hidroelektrik santrallerdedir (TEİAŞ, 2022).



Şekil 1.8 Gerçekleşen gerçek zamanlı enerji üretimi (TEİAŞ, 2022)

Bu çalışmanın birçok kurum ve kuruluşa vereceği katkılar aşağıdaki maddelerle açıklanmıştır.

- Son kullanıcı yani müşteri rüzgar türbinlerinin planlı bakımlarını tahminlenen üretime göre gerçekleştirebilir.
- Plansız bakımlarda ve arızalarda üretmesi gereken enerji miktarını önceden bilmesi oluşacak enerji kayıpları hesaplanabilir.

- Büyük parça değişimlerini planlarken rüzgara bağlı olarak üretimin az olacağı güne planlanarak olası enerji kayıplarını minimuma indirebilir.
- Benzer veri yapısına sahip araştırmacılar için bu çalışmada kullanılacak yöntemleri deneyerek kendi çalışmalarına katkı sağlayabilir.
- Rüzgar enerjisi sektörünün veri bilimi dünyası ile birleşimi sonucu endüstriye katkı sağlayabilir.

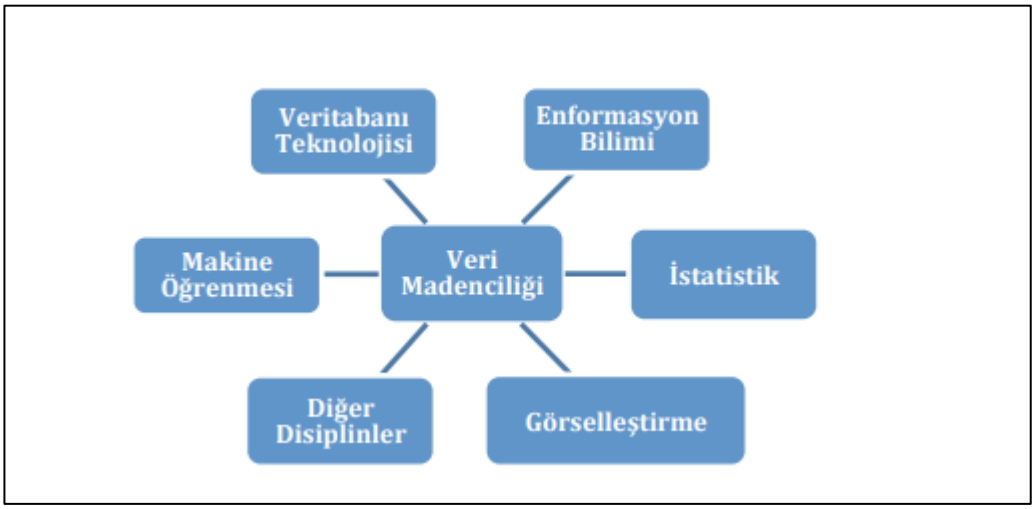
Bu tez çalışması 5 bölümden oluşmaktadır. Bölüm 1 bu çalışmanın arka planını, amacını ve önemini sunmakla birlikte rüzgar enerjisi ve çalışmada kullanılacak rüzgar türbini tipi hakkında bilgiler içermektedir. Dünyada ve Türkiye'de yenilenebilir kaynak olan rüzgar enerjisi hakkında güncel bilgiler verilmiştir. Bölüm 2'de veri madenciliği süreçleri hakkında genel bilgiler verilerek rüzgar gücü tahminleme çalışmaları hakkında literatür bilgisine yer verilmiştir. Bölüm 3'te bu çalışmada kullanılan makine öğrenmesi yöntemlerinin algoritmalarının işleyişi hakkında bilgi verilmiş ve detaylandırılmıştır. Aynı zamanda diğer algoritmalarla kıyaslama yapılarak avantaj ve dezavantajlarına değinilmiştir. Bölüm 4'te ise rüzgar gücünün tahminlenmesinde uygulanan adımlardan bahsedilerek veri analizini gerçekleştirmeden önce veri madenciliği süreci ile ilgili olarak veri ön işleme ve dönüştürme işlemlerinden bahsedilmiştir. Verinin tanımlayıcı ve keşifsel istatistikleri incelemiş ve veri yapısını anlamak adına bazı görsel analizler sunulmuştur. Ayrıca sonuçların değerlendirilmesinde kullanılan hata oranları hakkında da bilgi verilmiştir. Bölüm 5 sonuç bölümüne ayrılarak sonuçlar yorumlanmış ve kullanılan yöntemlerin karşılaştırılması ile uygun modele karar verilmesi hedeflenmiştir.

BÖLÜM 2
VERİ MADENCİLİĞİ SÜREÇLERİ VE İLGİLİ LİTERATÜR

Bu bölümde veri madenciliği tanımı ve süreçleri, makine öğrenmesinde kullanılan yöntemler ve literatürdeki çalışmalardan bahsedilmiştir.

2.1 Veri Madenciliği Tanımı ve Süreci

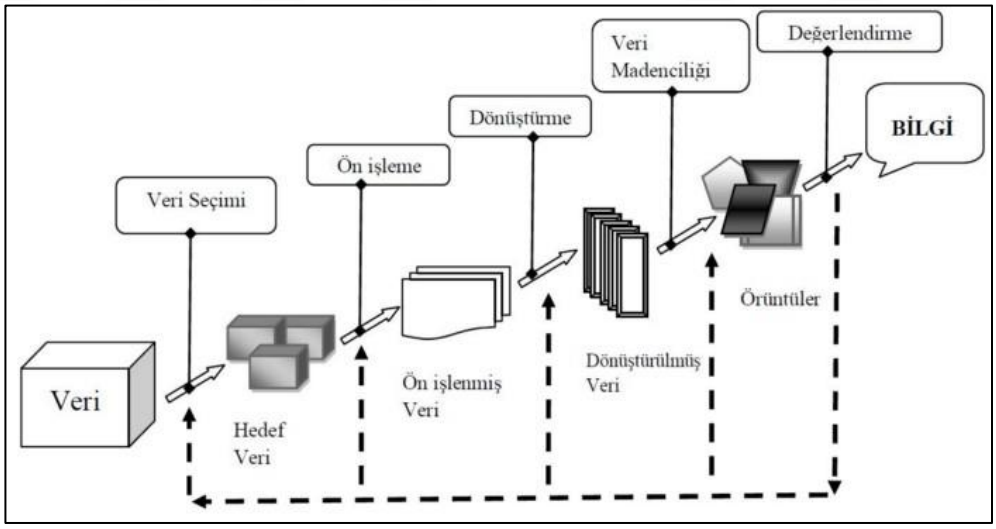
Veri madenciliği, beklenmeyen bağlantıları ortaya çıkarmak ve bilgileri veri sahibi için hem yararlı hem de anlaşılır yeni şekillerde sunmak için gözlemsel veri kümelerinin (tipik olarak büyük veriler) analizidir (Larose, 2005). Veri madenciliği, belirli bir amaç ile ilgili olan mevcut veri madenciliği yöntemlerini kullanma sürecidir. Bu adımın sonunda çeşitli örüntüler elde edilebilir. Veri madenciliği, büyük veritabanlarından bilgi çıkarma sorununu ele almak için makine öğrenimi, örüntü tanıma, istatistik, veritabanları ve görselleştirmeden elde edilen teknikleri bir araya getiren disiplinler arası bir alandır. Veri madenciliği Şekil 2.1'de görüldüğü gibi veri tabanı teknolojileri, enformasyon bilimi, görselleştirme, istatistik, makine öğrenmesi ve diğer disiplinleri içeren bir alandır.



Şekil 2.1 Veri madenciliğini oluşturan disiplinler (Özbay, 2015)

2.1.1 Veri Madenciliği Süreci

Bu çalışmada veri madenciliği sürecini işlerken rüzgar türbinlerinin gelişen teknoloji ile birlikte veri bilimi dünyasındaki işleyişini göz ardı etmemek gerekmektedir. Bununla birlikte, gelişen teknoloji ve artan kullanım alanının neticesinde rüzgar türbini bakımı da önemli bir konuma gelmiştir. Bakımlar, türbinlerin emre amade olmasını yani sürekli üretim yapılabilmesini sağlamak amacıyla yapılmaktadır ve işletme maliyetleri içerisinde önemli bir yere sahiptir. Bakım faaliyetleri kendi arasında planlı ve plansız bakım olarak alt kategorilere ayrılmaktadır. Plansız gelişen durumlarda bakımın maliyeti yüksek olabilir. Rüzgar türbinlerinin ne kadar üretim yapacağını önceden tahmin etmek planlı ve plansız bakımlar için kullanıcılara fikir verebilir. Bu çalışmada enerji üretim tahmin işleminde veri madenciliği yöntemi dikkate alınmıştır. Veri madenciliği çeşitli aşamalardan meydana gelmektedir. Şekil 2.2'de de görüleceği üzere bu aşamalar sırasıyla; veri seçimi, veri ön işleme, verinin dönüştürülmesi, veri madenciliği algoritmalarının uygulanması, değerlendirme aşamalarıdır.



Şekil 2.2 Veri madenciliği aşamaları (Taşcı, 2020)

Bu aşamaların her bir adımı ayrı ayrı önem içermektedir. Seçilen verinin veri madenciliğine uygun hale gelmesi için öncelikle bir ön işleme aşamasından geçmesi

gerekir. Bu aşamada veriler gürültülü verilerden temizlenir, eksik veriler tamamlanır (Taşcı, 2020).

Veri Seçimi: Bu aşamada önemli bir adım olan veri seçimi yapılır. Bu aşamadan sonra bu veriler üzerinde bir veri madenciliği süreci başlatılır.

Ön İşleme: Başarılı bir veri madenciliği çalışması, başlangıçta hazırlanan verilerin uygunluğuna ve kalitesine bağlıdır. Veriler ne kadar az gürültülü olursa, elde edilebilecek sonuçların kalitesi o kadar yüksek olur. Sorunlu verileri temizlemek, eksik verileri tamamlamak ve yeni nitelikler oluşturmak bu aşamanın temel unsurlarıdır.

Dönüştürme: Verilerin çeşitli veri madenciliği tekniklerine uygun formata dönüştürülmesi kısmıdır.

Veri Madenciliği: Mevcut veri madenciliği tekniklerinin belirtilen amaca uygun olarak kullanılması kısmıdır.

Yorumlama: Veri madenciliği kısmında belirlenen örüntülerin yeterli bilgiye sahip olup olmadığı yorumlanır. Bu inceleme sonucunda elde edilen örüntüler yetersiz bulunması takdirde önceki işlemler tekrarlanır (Akpınar, 2014).

2.2 Makine Öğrenmesinde Kullanılan Yöntemler

Veri madenciliğini oluşturan disiplinler arasında yer alan makine öğrenmesi yöntemini iki ana başlıkta tanımlayabiliriz bunlar; denetimli öğrenme ve denetimsiz öğrenme yöntemleridir.

2.2.1 Denetimli (Gözetimli) Öğrenme

Gözetimli makine öğrenimi olarak da bilinen denetimli öğrenme, makine öğrenimi ve yapay zekanın bir alt sınıflarından biridir. Verileri sınıflandırmak veya çıktıları doğru bir şekilde tahmin etmek için algoritmayı eğitmek üzere etiketli bir veri kümesi kullanılmaktadır. Denetimli öğrenme, bir modeli eğitmek ve istenen çıktıyı sağlamak için eğitim veri seti kullanmaktadır. Bu eğitim veri seti, modelin zamanla öğrenmesini sağlayan girdileri ve doğru çıktıları içermektedir. Algoritma, doğruluğunu bir kayıp fonksiyonu ile ölçer ve hata tamamen minimuma indirilene kadar ölçmeye devam eder.

Özetle, denetimli istatistiksel öğrenmede sınıflandırmayı ve tahminlemeyi sağlayacak bir model oluşturup, yeni veriler için bu model kullanılarak tahminler elde edilir. Veri madenciliği ile ilgili olarak, denetimli öğrenme iki tür problem olarak kategorize edilebilir; sınıflandırma ve regresyon.

- ***Sınıflandırma***: Test verilerini belirli kategorilerle doğru bir şekilde eşleştirmek için bu algoritma kullanır. Bir veri kümesindeki belirli girdileri tanır ve bu girdilerin nasıl etiketlenmesi veya tanımlanması gerektiği konusunda bazı sonuçlar çıkarmaya çalışır. Yaygın sınıflandırma algoritmaları arasında doğrusal sınıflandırıcılar, destek vektör makineleri (SVM'ler), karar ağaçları, k-en yakın komşu ve rastgele ormanlar yer almaktadır.

- ***Regresyon:*** Bağımlı ve bağımsız değişkenler arasındaki ilişkiyi anlamak için kullanılır. Diğer deyişle, regresyon analizi bağımsız değişkenin aldığı değere bağlı olarak bağımlı değişkenin değerini tahmin etmek ve bağımsız değişkendeki değişimin bağımlı değişkenin ortalaması üzerindeki etkisini açıklamak üzere kullanılır. Doğrusal regresyon, lojistik regresyon ve polinom regresyon yaygın regresyon algoritmalarıdır.

Denetimli makine öğrenme süreçlerinde çeşitli algoritmalar ve hesaplama teknikleri kullanılmaktadır. Genellikle R ve Python gibi programlama dilleri kullanılarak hesaplanan en yaygın öğrenme yöntemlerinden bazıları şunlardır: Yapay sinir ağları *(neural network*), Naïve Bayes sınıflandırma yöntemi, lineer regresyon (*lineer regression*), lojistik regresyon (logistic regression), destek vektör makinesi (*support vector machine*), k-en yakın komşuluk (*k-nearest neighbor*), rassal ormanlar (*random forest*). Denetimli öğrenme modelleri, çeşitli iş uygulamalarını oluşturmak ve geliştirmek için kullanılabilir. Bunlar arasında görüntü ve nesne tanıma olarak bilinen denetimli öğrenme algoritmaları, videolar ve görüntülerdeki nesneleri tanımlamak, ayırmak ve sınıflandırmak için kullanılmaktadır. Bu durum onları çeşitli bilgisayarla görme teknikleri ve görüntü analizi uygulamaları için kullanışlı duruma getirmektedir. Çeşitli veri noktalarına ilişkin detaylı bilgiler elde etmek için denetimli öğrenme modelleri yaygın olarak kullanılmaktadır. Bu modeller yardımı ile girdi değişkenleri kullanılarak çıktı değişkeni tahmin edilebilir ve elde edilen sonuçlar süreçler hakkında planlama ve karar alma süreçlerini güçlendirir. Müşteri duyarlılığı

analizinde ise şirketler, denetimli makine öğrenimi algoritmalarından faydalanarak, az insan müdahalesiyle büyük veri kümelerinden bağlam, duygu ve amaç gibi önemli bilgileri çıkarabilir ve sınıflandırabilir. Bu durum, müşteri bağlantılarını iyi anlamamıza yardımcı olabilmektedir. İstenmeyen posta algılama yani spam tespiti olarak bilinen diğer bir uygulama ise denetimli öğrenme modelinin başka bir örneğidir. Firmalar, gözetimli sınıflandırma algoritmalarından faydalanarak yeni verilerdeki kalıpları veya anormallikleri tanımak ve spam olan ve olmayan iletişimleri etkili bir şekilde organize etmek için veritabanlarını eğitebilir (IBM Cloud Education, 2020). Denetimli öğrenmenin de bazı dezavantajları vardır. Bunlar;

- Denetimli öğrenme modelinin eğitimi uzun zaman alabilir.
- Veri kümelerinin, algoritmaların yanlış öğrenmesine neden olan insan hatasına maruz kalması muhtemeldir.
- Denetimsiz öğrenme yöntemlerinden farklı olarak denetimli öğrenme, veriyi kendisi gruplandıramaz veya sınıflandıramaz.

2.2.2 Denetimsiz (Gözetimsiz) Öğrenme

Denetimsiz makine öğrenimi algoritmaları hiçbir etiket verilmeden veri kümelerini analiz etmek için kullanılmaktadır. Bu algoritmalar, insan müdahalesi olmadan verilerdeki gizli kalıpları keşfetmektedir. Bilgilerdeki benzerlikleri ve farklılıkları belirleme yeteneği, denetimsiz makine öğrenimini keşifsel veri analizi, çapraz satış stratejileri, müşteri segmentasyonu ve görüntü tanıma için ideal bir çözüm haline getirmektedir. Denetimsiz öğrenme yöntemlerinde, kümeleme, boyut azaltma ve birliktelik kuralları gibi yöntemler önem arz etmektedir.

- ***Kümeleme:*** Etiketlenmemiş verilerin benzerlik veya farklılıklarını gruplandırır. Kümeleme algoritmaları özetlemek veya verileri yapılandırmak için kullanılır. Kümeleme algoritmaları arasında özel, örtüşen, hiyerarşik ve olasılıksal olmak üzere çeşitli türleri bulunmaktadır.

- ***Birliktelik kuralları:*** Belirli bir veri kümesinde değişkenler arasındaki ilişkileri bulmanın kural tabanlı bir yöntemidir. Pazar sepeti analizi için yaygın olarak kullanılan bu yöntemler, kuruluşların farklı malzemeler (ürünler) arasındaki ilişkiyi anlamalarını sağlar. İşletmeler, tüketici alışkanlıklarını anlayarak çapraz satış stratejileri ve öneri arama motorları geliştirebilir. Bunun örnekleri Amazon'un "Bu Ürünü Satın Alan Müşteriler Ayrıca Satın Aldı" veya Spotify'ın "Haftalık Keşfet" çalma listesinde görülebilir. *Apriori, Eclat* ve *FP-Growth* gibi birliktelik kurallarını oluşturmak için kullanılan farklı algoritmalar olsa da, en yaygın olarak Apriori algoritması kullanılır.

- ***Boyut azaltma:*** Büyük veri kümeleri genellikle daha doğru sonuçlar verir, ancak makine öğrenimi algoritmalarının performansını etkileyebilir ve veri kümesinin görselleştirilmesini zorlaştırabilir. Boyut azaltma, belirli bir veri setinde çok fazla özellik veya boyut olduğunda kullanılan bir tekniktir. Veri girişi sayısını yönetilebilir bir boyuta indirgerken, veri kümesi bütünlüğünü de mümkün olduğunca korumaktadır (IBM Cloud Education, 2020).

2.3 Literatür Taraması

Literatürde rüzgar gücü tahminleri için birçok makine öğrenmesi yöntemlerinden yararlanılmaktadır.

Singh vd. (2021) tarafından Türkiye'nin batısında yer alan bir rüzgar santralinde rüzgar hızı ve yönüne ilişkin tarihi verilere dayalı enerji üretimini kısa süreli tahmin doğruluğunu iyileştirmek için beş farklı yöntemi karşılaştırmışlardır. Bu yöntemler; rassal ormanlar (*Random Forest – RF*), gradyan arttırma regresyon ağaçları (*Gradient Boosting Regression – GBR*), k en yakın komşu (*K-Nearest Neighbor – k-NN*), karar ağacı (*Decision Tree – DT*) ve ekstra ağaç regresyonu (*Extra Tree Regression – ET*) şeklindedir. Çalışmada SCADA'dan elde edilen bir yıllık 10 dakikalık geçmiş rüzgar hızı, rüzgar yönü, aktif güç ve teorik güç verileri kullanılmıştır. Rüzgar türbinin tipine bağlı olarak başlatma rüzgar hızı olan 3 m/s, negatif ve üretim olmayan verileri veri setinden çıkararak ön işleme yapılmış olup aynı zamanda verileri maksimum-

minimum normalizasyonu kullanılarak normalleştirilmiştir. Veri ön işlemeden sonra makine öğrenmesi algoritmalarını uygulayarak hata ölçüleri olan ortalama mutlak hata (*Mean Absolute Error - MAE*), ortlama mutlak yüzde hatası (*Mean Absolte Percentage Error – MAPE*), kök ortalama kare mutlak (*Root Mean Square Error – RMSE*) ve belirtme katsayısı (R^2) yöntemleriyle kıyaslayarak en iyi performansın gradyan arttırma regresyon ağaçları (GBR) olduğunu saptamışlardır.

Yürek vd. (2021) tarafından geçmiş rüzgar enerjisi üretim verileri ve hava durumu tahmin raporlarını kullanarak yenilenebilir enerji sektöründeki gerçek yaşam sorununu, önümüzdeki 24 saat için saat başı rüzgar enerjisi üretim miktarını makine öğrenmesi (ML) teknikleri ile tahmin edebilmek için bir model önermişlerdir. Önerilen yaklaşım öncelikle, veri setini anlamlı kümeler halinde gruplamak için denetimsiz bir ML yöntemi olan k-ortalama kümeleme algoritması kullanılır. Daha sonra, bu kümeler yeni öznitelik değerleri olarak kabul edilir ve son olarak, tahmin için denetimli bir ML yöntemi (regresyon) gerçekleştirilir. Bu çalışma dokuz denetimli öğrenme algoritmasını karşılaştırmaktadır: K-En Yakın Komşu, Destek Vektör Regresyonu, Rassal Orman, Ekstra Ağaçlar, Gradyan Artırma, Ridge Regresyon, En Küçük Mutlak Daralma ve Seçme Operatörü, Karar Ağacı, ve Konvolüsyonel Sinir Ağı. Bu yöntemler arasında en düşük hata ölçüsüne sahip olan model Rassal Orman algoritması çıkmıştır.

Alkesaiberi vd. (2022) tarafından Fransa, Türkiye ve Kaggle'daki üç rüzgar türbininden alınan gerçek ölçümleri kullanarak makine öğrenmesi methodlarını karşılaştırmıştır. Çalışmada metodoloji olarak farklı çekirdekler ile Gauss süreç regresyonu (*Gaussian processes regression - GPR*) ve destek vektörü regresyonu (SVR), topluluk öğrenme (*Ensemble learning - EL*) modelleri (yani rassal orman vb.) ve optimize edilmiş GPR ve EL modelleri rüzgar gücünü tahmin etmek için karşılaştırılmıştır. İlk olarak, makine öğrenimi modelini oluşturmadan önce toplanan verileri ön işleme tabi tutulmaktadır. Tam olarak, veri kalitesini iyileştirmek için aykırı gözlemler için çeşitli işlemler yapılmıştır. Rüzgar türbini ölçümlerindeki aykırı değerler, arızalı ölçüm sensörleri gibi çok çeşitli nedenlere bağlı olabilir. Aykırı değerler genellikle belirlenir ve dikkate alınan modelin tahmin doğruluğunu artırmak

için elimine edilir. Aksi takdirde, oluşturulan model yanlı veya hatalı olabilir. Burada aykırı değerler, eğitim verilerinin medyanı ile değiştirilmiştir. Rüzgar hızındaki eksik değerler, yanlış veri kaydı, gök gürültülü fırtınalar, arızalar ve diğer anemometre arızaları gibi farklı nedenlerden kaynaklanabilir. Literatürde kayıp değer atama için çeşitli teknikler geliştirilmiştir. Bu çalışmada, kayıp değer ataması için R paketi *Amelia* kullanılmıştır. Ardından, normalleştirilmiş verileri eğitim ve test kümelerine ayrılmıştır. Modeller öncelikle eğitim seti bazında eğitilir ve model parametrelerinin değerleri hesaplanır. Burada, makine öğrenimi modellerinin optimal parametrelerini bulmak için Bayes optimizasyonu uygulanmıştır. K-kat çapraz doğrulama tekniği, bu modellerde tavsiye edildiği gibi eğitim verilerine dayalı olarak oluşturulurken dikkate alınmıştır. Özellikle, araştırılan modellerin eğitiminde beş katlı çapraz doğrulama tekniği uygulanmıştır. Son olarak, her modelin performansı için üç istatistiksel hata fonksiyonu (RMSE, MAE ve R^2) kullanılarak değerlendirilmiştir. Sonuç olarak rüzgar enerjisi serilerinin geçmiş gözlemlerinden ilgili bilgileri dikkate alan dinamik modellerin kullanılmasının, statik modellere kıyasla tahmin sonuçlarını iyileştirdiğini ortaya koymuştur. Tüm modeller için diğerlerine hakim olan tek bir yaklaşım olmadığı gözlemlenmiştir, ancak ortalama olarak, GPR modelleri ve topluluk modelleri, yaklaşık 0,95'lik bir R^2 ile tatmin edici bir tahmin performansına ulaşmıştır.

Fugon vd. (2008) rüzgar enerjisi tahmini için veri madenciliği türündeki modelleri değerlendirmişlerdir. İncelenen modeller arasında; sinir ağları, destek vektör makineleri, yakın zamanda önerilen regresyon ağaçları yaklaşımlarıdır. Üç farklı gerçek rüzgar santrali için değerlendirme sonuçları sunulmaktadır. Fransa'da WF1, WF2 ve WF3 olarak belirtilen üç rüzgar santrali değerlendirilmektedir. Bu santraller çeşitli arazi ve iklim koşullarına sahiplerdir. Saatlik elektrik üretim zaman serileri Temmuz 2004'ten Aralık 2005'e kadar 18 aylık bir süreyi kapsamaktadır. Aynı dönem için Meteo France'ın ARPEGE modeline göre sayısal hava durumu tahminleri (*Numerical Weather Prediction - NWP*) kullanılmıştır. Tahminler, 0 ile 60 saat ufku için 3 saatlik bir çözünürlükle günde bir kez sağlanmıştır, yani her bir meteorolojik değişken için çalışma başına 20 değer sağlanmıştır. Bu çalışmada dikkate alınan meteorolojik değişkenler, yer seviyesinden 50 metre yükseklikteki rüzgar hızı ve şiddetli rüzgar yönüdür. Çalışmada performansları kıyaslamak üzere

Normalleştirilmiş Kök Ortalama Kare Hatası (*Normalized Root Mean Square Error – NMAE*) ve Normalleştirilmiş Ortalama Mutlak Hata (*Normalized Mean Absolute Error – NMAE*) metotları kullanılmıştır. Buna göre en başarılı performansa sahip olan algoritmanın rassal ormanlar olduğu gözlemlenmiştir.

Buturache ve Stancu (2021) yapay sinir ağları, destek vektör regresyonu, rastgele ağaçlar ve rastgele orman konularında karşılaştırmalı ve kapsamlı bir çalışma önermiştir ve yukarıda bahsedilen tekniklerin uygulanmasının artılarını ve eksilerini sunmuşlardır. *CRISP-DM* olarak bilinen veri madenciliği çerçevesine dayanan adım adım bir yaklaşım ile metrik seçimi, model seçimi veya hiperparametre ayarlama dahil olmak üzere süreci uçtan uca değerlendirmişlerdir. Bu çalışma, üretimde uygulanmayacak, ancak teorik hususlar ve çalışmanın sonuçları, rüzgar türbini endüstrisindeki gerçek operasyonları desteklemek için uygulanabilecek teknik bir çözümün geliştirilmesine yönelik alanı hazırlamayı amaçlamışlardır. Açık kaynak bir veri seti kullanarak makine öğrenmesi algoritmalarını karşılaştırmışlardır. Bu çalışmada enerji, rüzgar hızı (2 ve 10 metrelik ölçümler), hava yoğunluğu, sıcaklık, hava basıncı gibi değişkenler kullanılmıştır. Kullanılan yöntemler ise yapay sinir ağları (A*rtficial Neural Network – ANN),* destek vektör makineleri (*Support Vector Regression – SVM*), regresyon ağaçları ve rassal ormanlar (RF) şeklindedir. Rüzgar enerjisini tahmin etme kapasitesini değerlendirirken ele alınan kıyaslama metrikleri, ilk olarak belirtme katsayısı olarak bilinen R^2 model tarafından açıklanan bağımlı değişkendeki varyans yüzdesidir. RMSE ve MAE aynı veri setindeki farklı modeller arasında karşılaştırmalar yapılmasını sağlayan ölçeğe bağlı doğruluk ölçüleridir. Dördüncü metrik ise eğitim süresi ile temsil edilir yani süre bakımından makine öğrenmesi algoritmalarının performansının değerlendirildiği kısımdır. Optimize edilmiş en nihai parametreler belirlendikten sonra yapay sinir ağları diğer üç algoritmaya göre daha iyi performans sergilediği gözükmektedir. Parametre seçimindeki hızlı performansı ile diğer yöntemlerin önüne geçmiştir.

Eyecioğlu vd. (2019) rüzgar türbini güç üretiminin tahmini için doğrusal regresyon, k-en yakın komşu regresyonu ve karar ağacı regresyonu olarak adlandırılan üç makine öğrenme algoritmasının performansı karşılaştırmışlardır. Tahmin aşamasında rüzgar

hızı, rüzgar yönü, barometrik basınç ve hava sıcaklığı gibi meteorolojik parametrelerden yararlanılmıştır ve bunların rüzgar gücü parametresi üzerindeki önem faktörleri de ortaya çıkmaktadır. Bu çalışmada yapılan genel analizler sonucunda rüzgar enerjisi tahminleri doğru bir şekilde elde edilmektedir. Kullanılan veri seti Danimarka Teknik Üniversitesi'nden açık kaynaktan alınmıştır. 10 dakikalık aralıklarla elde edilen 4464 gözlemden oluşmaktadır. Değişkenler; rüzgar hızı, rüzgar yönü, barometrik basınç, hava sıcaklığı ve rüzgar gücü şeklindedir. Nokta grafiğine bakılarak her bir parametrenin korelasyonlarına bakılarak rüzgar gücü ile pozitif korelasyona sahip rüzgar hızı çıkmıştır. Bahsedilen makine öğrenmesi algoritmalarının doğruluk oranları k-kat çapraz geçerlilik ile k'yı 10 seçerek tahmin edilmiştir. Veri seti %80 eğitim ve %20 test olacak şekilde ayrılmıştır. Bu çalışmada, rüzgar türbini güç üretimi üzerindeki tahmin performanslarını karşılaştırmak için bir yandan lineer regresyon, k-en yakın komşu regresyonu ve karar ağacı regresyon algoritmaları kullanılmıştır. Ortalama mutlak hata ve belirtme katsayısının hata ölçüleri açısından, en yakın komşu regresyon ve karar ağacı regresyon algoritmaları, lineer regresyon algoritmasından daha iyi tahmin performansı sağladığı görülmüştür. Rüzgar hızı, rüzgar yönü, barometrik basınç ve hava sıcaklığı parametreleri ise rüzgar türbini güç üretimini etkileyen meteorolojik parametreler olarak kabul edilmektedir. Elde edilen önem faktörü açısından rüzgar hızı parametresi, barometrik basınç, rüzgar yönü ve hava sıcaklığı parametrelerinden daha umut verici bir etki göstermektedir. Gelecekteki çalışmalarda, daha derin analizler için daha fazla makine öğrenmesi algoritması ve daha fazla meteorolojik parametre dikkate alındığını ortaya koymuşlardır.

Son vd. (2019) çalışmasında rüzgar enerjisini tahmin etme yöntemlerini dört kategoriye ayırmaktadır: fiziksel yöntemler, istatistiksel yöntemler, yapay zeka yöntemleri ve hibrit yöntemler. Bu çalışmada, kısa vadeli rüzgar gücünü tahmin etmek için modifiye LSTM (*Long/Short Term Memory* – Uzun/Kısa Süreli Bellek) kullanan bir hibrit model önermişlerdir. Değiştirilmiş LSTM tarafından benimsenen veriler, rüzgar gücünün tahmin faktörleri olan önceki veriler yerine mevcut gözlem verilerini (rüzgar gücü, rüzgar yönü ve rüzgar hızı) kullanmaktadır. Değiştirilmiş LSTM'nin performansı, mevcut gözlem verilerinin birleştirilmesinden türetilen dört çok

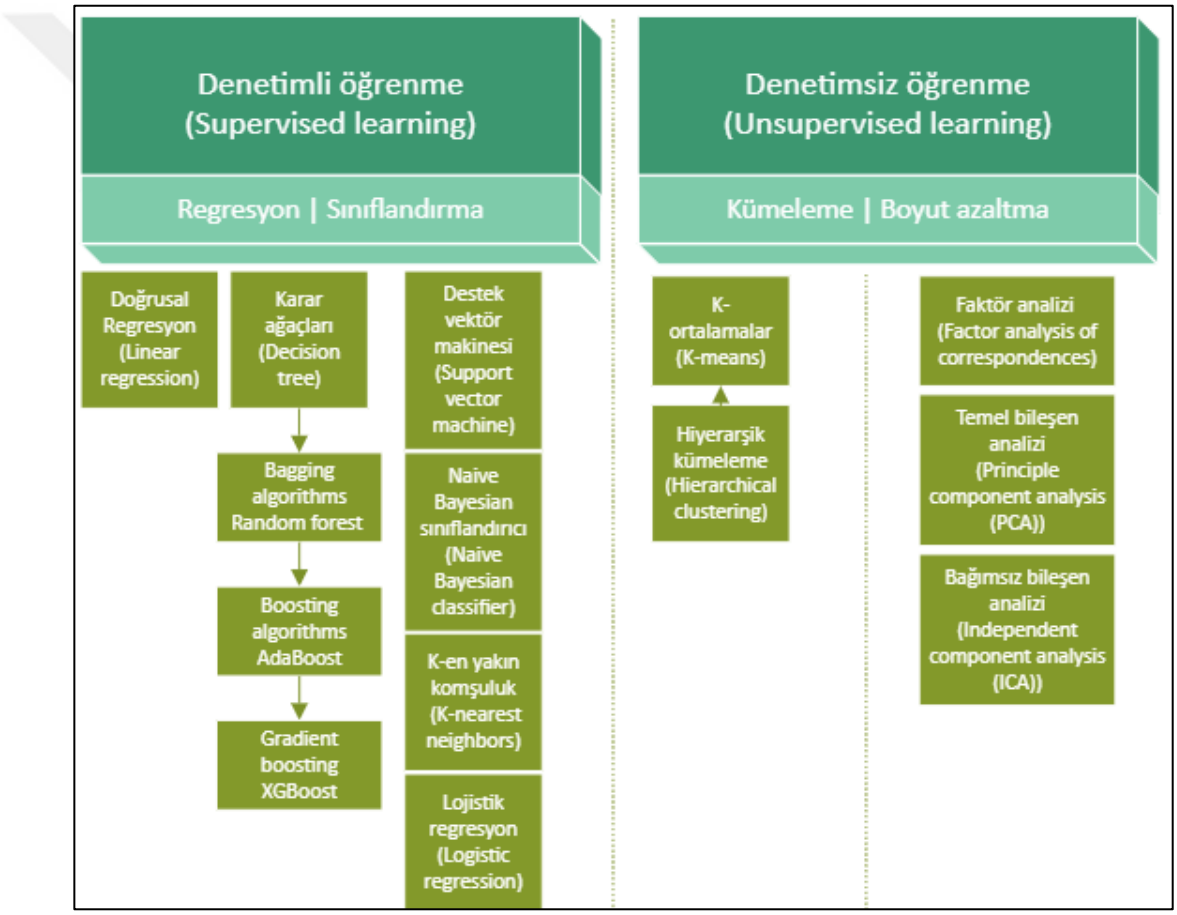
değişkenli model arasında karşılaştırılmıştır. Çok değişkenli modeller arasında, önerilen hibrit yöntem, Model 1 (rüzgar gücü) ile ilk aşamada iyi performans gösterirken, Model 3 (rüzgar gücü, rüzgar hızı) ile orta ve son aşamalarda kısa vadeli rüzgar gücü tahmininde mükemmel performans göstermiştir. Deney sonuçları, önerilen modelin kısa vadeli rüzgar enerjisi tahmininde diğer modellere göre daha sağlam ve doğru olduğunu göstermiştir. Tablo 2.1'de rüzgar gücü tahminlemelerinde kullanılan ve geliştirilen bazı çalışmaların özetini bulabilirsiniz.

Tablo 2.1 Son beş yıl içinde geliştirilen bazı çalışmalar

Referans	**Model**	**Veri seti bilgisi**	**Hata fonksiyonu**
Shabbir vd. (2021)	Destek Vektör Makinesi (SVM), Doğrusal Olmayan Otoregresif Sinir Ağları (NAR) ve Tekrarlayan Sinir Ağı-Uzun Süreli Kısa Süreli Bellek (RNN-LSTM)	Gerçekleşen enerji üretimi	RMSE
Huang ve Jiang (2022)	DNN(deep neural network), LSTM, Informer, MSIN (Multi-step Informer network)	Rüzgar hızı, hava basıncı, rüzgar enerjisi	RMSE, MAE, MSE
Singh ve Rizwan (2022)	RF, LSTM, XGBoost	Aktif güç (kW), rüzgar hızı (m/s), teorik güç (kWh), rüzgar yönü(º)	MSE, RMSE, MAPE, MAE, R^2

BÖLÜM 3
KULLANILAN MAKİNE ÖĞRENMESİ YÖNTEMLERİ

Bu bölümde analizde kullanılan makine öğrenmesi algoritmalarının tanımları, algoritma adımları ve kullanılmasındaki avantajları verilmiştir. Aynı zamanda göstermiş olduğu performansa bağlı olarak diğer algoritmalarla kıyaslama yapılmıştır. Bu yöntemlerle ilgili halihazırda sürekli geliştirmeler yapılmakta ve makine öğrenimi dünyasına katkılar yapılmaya devam edilmektedir. Şekil 3.1'deki sınıflandırma diyagramında bazı makine öğrenmesi yapıları gösterilmiştir.



Şekil 3.1 Makine öğrenmesi algoritmalarının sınıflandırma diyagramı

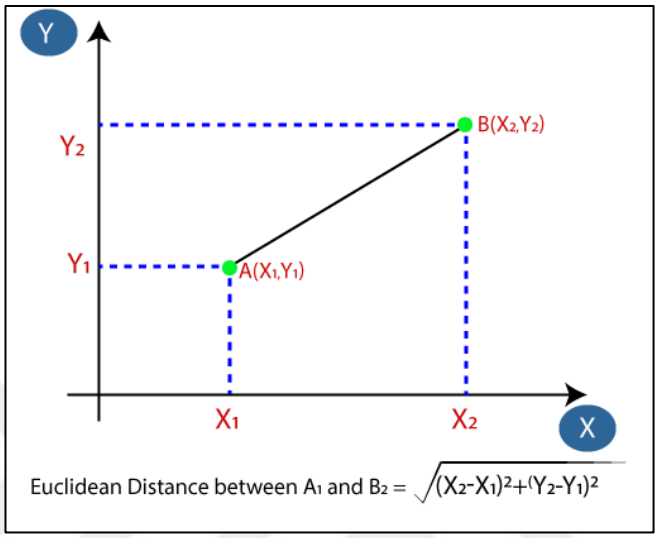
3.1 K-en Yakın Komşuluk Regresyonu (KNN)

K-en yakın komşuluk algoritması, Cover ve Hart (1967) "*Nearest neighbor pattern classification* (En yakın komşu örüntü sınıflandırması)" tarafından ve Cover (1968) "*Rates of Convergence for Nearest Neighbor Procedures* (En Yakın Komşu Prosedürleri için Yakınsama Oranları)" tarafından önerilmiştir. Bu algoritmda test örneği ile tüm örnekler arasındaki mesafe hesaplanır ve karar olarak en yakın komşunun sınıfı kullanılacak şekilde adımlar uygulanır.

K-en yakın komşuluk (k-NN), parametrik olmayan bir sınıflandırma algoritmasıdır, yani temel veri kümesi üzerinde herhangi bir varsayımda bulunmaz. Sadeliği ve etkinliği ile bilinir. Sınıflandırmada, farklı özellikler, etiketlenmemiş verilerin ait olduğu sınıfı belirler. k-NN çoğunlukla sınıflandırıcı olarak kullanılır. Belirli bir bölgedeki en yakın veya komşu eğitim örneklerine dayalı olarak verileri sınıflandırmak için kullanılır. Bu yöntem kolay ve düşük hesaplama süresine sahiptir. Sürekli veriler için en yakın komşularını hesaplamak için Öklid mesafesini kullanır. Yeni bir girdi için *K* en yakın komşu hesaplanır ve komşu verilerin çoğunluğu yeni girdi için sınıflandırmaya karar verir. Hesaplama maliyeti biraz yüksektir, çünkü tüm hesaplamalar veri setinde karşılaşıldığında değil, eğitim verileri sınıflandırılırken yapılır. Veri kümesi eğitilirken, eğitim verilerini depolamak ve bunun yerine veri kümesini ezberlemek dışında pek bir şey yapılmadığı için tembel bir öğrenme algoritmasıdır. Eğitim veri seti üzerinde genelleme yapmamaktadır. Bu nedenle, test aşamasındayken eğitilen tüm temel veri kümesi gereklidir.

Bir regresyon probleminde k-NN algoritması sürekli bir değişkene ait değeri tahmin eder. Bu değer, *K* en yakın komşularının değerlerinin ortalamasıdır (Taunk vd., 2019). k-NN regresyonu, aynı komşuluktaki gözlemlerin ortalamasını alarak bağımsız değişkenler ve sürekli sonuç arasındaki ilişkiyi sezgisel bir şekilde tahmin eden parametrik olmayan bir yöntemdir. Komşuluk boyutunun analist tarafından ayarlanması gerekir veya ortalama kare hatasını en aza indiren boyutu seçmek için çapraz doğrulama kullanılarak seçilebilir (Teixeira-Pinto, 2022). Kullanılan yaygın benzerlik ölçüsü Şekil 3.2'deki gibi Öklid mesafesidir. Alternatif olarak, soruna bağlı

olarak *Manhattan, Minkowski* ve *Hamming* uzaklık ölçüleri de kullanılabilir. Uzaklık hesaplandıktan sonra sıralanır ve gelen değer uygun olan sınıfa atanır. Mesafe ölçüsünün de yalnızca sürekli değişkenler için geçerli olduğu unutulmamalıdır. Kategorik değişkenler durumunda *Hamming* mesafesi kullanılmalıdır.



Şekil 3.2 Öklid uzaklık hesabı ve gösterimi

K-en yakın komşuluk (k-NN) modeli tüm eğitim veri setini saklar ve yeni bir test veri noktası verildiğinde amaç, eğitim setinden test veri noktasına en yakın olan *k* noktaları bulmaktır. Yani burada benzerlik metriğini en aza indiren noktaları bulmaya çalışmaktadır. Özellikle standartlaştırma, k-NN gibi algoritmalarda önemli bir adımdır çünkü burada veri kümesinin dağılımına bağlı olan Öklid mesafesi gibi ölçümlerle uğraşılmaktadır. Bu yüzden uygun bir model oluşturmak için veri setini standartlaştırmak gerekmektedir. k-NN'nin temel avantajları arasında kolay yorumlanabilirlik, sınıflandırma ve regresyon problemlerinde kullanım için çok yönlülük ve parametrik olmayan doğası bulunmaktadır (herhangi bir varsayıma veya veri ayarlamasına gerek yoktur). k-NN algoritmasının temeli üç ana adımda özetlenebilir:

- Test veri seti ile eğitim veri seti arasındaki önceden tanımlanmış mesafe hesaplanır.

$$\text{Öklid} = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2} \quad (3.2)$$

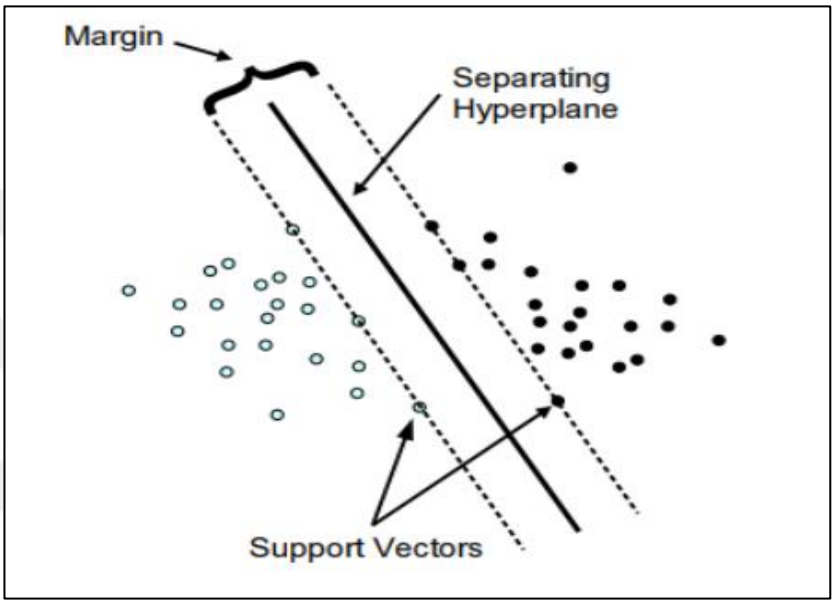
- Eğitim veri setinden k-minimum mesafeli k-en yakın komşular seçilir.
- Ağırlıklı ortalama yaklaşımına dayalı olarak nihai yenilenebilir enerji çıktısı tahmin edilir (Vontela, 2022).

"*K*" giriş değişkeni, dikkate alınması gereken komşuların sayısına karar verir. "*K*" değeri, algoritmayı her sınıfın sınırlarını oluşturabileceğimiz "*K*" değerini kullanarak algoritmayı etkiler. En iyi *K* değeri için, önce veriler incelenerek seçilir. Daha büyük *K* değerleri, gürültüyü azalttığı için daha kesin sonuçlar verir ancak bu garanti bir çözüm değildir. İyi bir *K* değeri, çapraz doğrulama kullanılarak da belirlenebilir. *K*=1 ise, veri basitçe en yakın komşusunun sınıfına atanır. *K*=1 olduğunda eğitim verileri için hata oranı sürekli olarak sıfır çıkacaktır. Bunun nedeni, herhangi bir eğitim veri noktasına en yakın noktanın kendisi olmasıdır. Dolayısıyla en iyi sonuçlar *K*=1 ise elde edilir. Ancak *K*=1 ile aşırı uyum sorunu ortaya çıkmaktadır. Uygun bir *K* değeri elde etmek için eğitim ve doğrulama seti, ilk veri setinden ayrılmalıdır. En yakın iki komşu (*K*=2) iki farklı sınıfa aitse sonuç bilinmemekle birlikte daha büyük bir değere yükseltilirse (en yakın 5 komşu) bu, en yakın komşu bölgeyi tanımlayacak ve netlik sağlayacaktır. Daha büyük "*K*" değerleri, sınıf sınırlarını daha pürüzsüz hale getirir. Bu durum diğer sınıfların noktaları aynı kümeye dahil edilebileceği için tercih edilmeyebilir. Eğitim veri noktaları dağınık bir şekilde mevcut olduğunda, K değerini belirlemek zordur (Taunk, De, Verma ve Swetapadma, 2019).

3.2 Destek Vektör Regresyonu (SVR)

Destek vektör makinesi (SVM), verileri farklı sınıflara ayırmak için kullanılan denetimli bir makine öğrenme algoritmasıdır. Çoğu algoritmanın aksine, SVM, çeşitli sınıflar arasında bir karar sınırı gibi davranan bir hiper düzlem kullanır. SVM'lerin çeşitli ortamlarda iyi performans gösterdiği bilinmektedir ve genellikle en iyi sınıflandırıcılardan biri olarak kabul edilir. SVM, etiketli (label) eğitim verilerini inceler ve ardından eğitim aşamasında öğrendiklerine bağlı olarak yeni gözlem

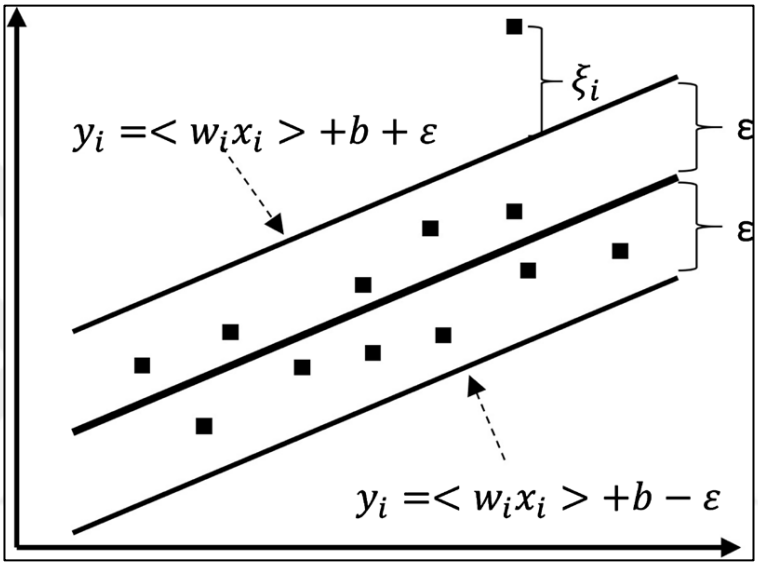
verilerini sınıflandırır. SVM'nin temel bir avantajı, hem sınıflandırma hem de regresyon problemleri için kullanılabilmesidir. SVM esas olarak sınıflandırma için bilinmesine rağmen, destek vektör regresyonu (SVR) regresyon problemleri için kullanılır. Tarihte Bernhard Boser, Isabelle Guyon ve Vladimir Vapnik doğrusal olmayan sınıflandırıcılar için "kernel trick (kernel hilesi)" önererek maksimum marjlı hiper düzlemler oluşturmuşlardır.



Şekil 3.3 Hiper düzlem ve destek vektörleri (Agarwal, 2021)

Hiper düzlemler, veri noktalarının sınıflandırılmasına yardımcı olan karar sınırlarıdır. Şekil 3.3'te görüldüğü gibi alt düzlemin her iki tarafına düşen veri noktaları farklı sınıflara ayrılır. Ayrıca, hiper düzlemin boyutu özelliklerin sayısına bağlıdır. *p* boyutlu bir uzayda hiper düzlem boyutu (*p-1*) kadardır. Yani, iki boyutlu bir uzayda bir düzlem bulunmaktadır. Hiper düzlem ile eğitim setindeki herhangi bir nokta arasında mümkün olan en yüksek marja sahip bir hiper düzlem seçmek ve yeni verilerin doğru bir şekilde sınıflandırılma şansını arttırmaktır. Destek vektör regresyonunda amaç, bir marjın aralığına maksimum noktayı en küçük hata ile alabilecek şekilde doğru ya da eğri (fonksiyon) belirlemektir. Aykırı gözlemlerin ortaya koyacak olan modeli belirlemede büyük bir önemi vardır. SVR dayanıklıdır. Aykırı gözlemlerin regresyon denklemi üzerindeki etkilerini minimize eder. Eşik değer ile gözlemlerin regresyon eğrisine olan katkıları ayarlanabilmektedir. Doğrusal

regresyon veya ileri beslemeli sinir ağı, SVR'deki hatayı en aza indirmeyi amaçlarken, amaç hatayı sabit bir eşik içine sığdırmaktır. Bu şekilde, SVR, doğru karar sınırını seçmedeki problemlerle ilişkilendirilir. En iyi uyum, sınırlar arasına maksimum sayıda veri noktası eklendiğinde elde edilir. Karar sınırının nereye yerleştirileceğine karar vermek ve hiperdüzlemden ε ve −ε uzaklıklarını hiperdüzleme en yakın veri kayıtlarının Şekil 3.4'te görsel olarak temsil edilen sınırların içinde olacağı şekilde ayarlamak için çaba harcanır (Buturache ve Stancu, 2021).



Şekil 3.4 Destek vektör regresyon grafiği (Buturache ve Stancu, 2021)

Eğitim ve test alt kümelerine ayrılmış belirli bir veri seti için, eğitim noktaları (x_1, y_1), (x_2, y_2), … , $(x_i, y_i) \subset XxR$, burada X girdi uzayı olarak R^d ve i=1,2, …, n kapsar. $f(x)$ fonksiyonu hiperdüzlemden en büyük ε sapmasına sahip olmalı ve mümkün olduğu kadar düz olmalıdır. Lineer bir fonksiyon için:

$$f(x) = \langle w, x \rangle + b \tag{3.3}$$

burada, $w \in X$, $b \in R$ ve $\langle ., . \rangle$ X'deki nokta çarpımını temsil eder.

Düzleştirilmiş bir $f(x)$ aramak, w'nin değerinin en aza indirilmesi gerektiği anlamına gelir.

Burada:

$$minimize \frac{1}{2}\|w\|^2 \tag{3.4}$$

ilgili kısıtlar ise,

$$y_i - \langle w_i, x_i \rangle - b \leq \varepsilon \tag{3.5}$$

$$y_i - \langle w_i, x_i \rangle + b \leq \varepsilon \tag{3.6}$$

Ancak bu ideal kısıtlardır. Sınırları aşan -yani ε'dan büyük hatalar- aylak değişkenler S_i ve S_i^* optimizasyona dahil edilir.

$$Minimize \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{l}(S_i + S_i^*) \tag{3.7}$$

yeni ilgili kısıtlamalar ile,

$$y_i - \langle w_i, x_i \rangle - b \leq \varepsilon + S_i \tag{3.8}$$

$$y_i - \langle w_i, x_i \rangle + b \leq \varepsilon + S_i^* \tag{3.9}$$

burada, S_i, $S_i^* \geq 0$.

Yeni hiperparametre *C*, *f*'in düzlüğünü etkileyerek sınırları aşan veri noktalarındaki toleransı azaltabilir veya artırabilir. *C*'nin değeri ne kadar büyük olursa ε'dan büyük sapmalara karşı tolerans o kadar büyük olur. Lineer olarak ayrılamayan özellikler için, onları lineer olarak ayrılabilir yapmak için bir seçenek, onları lineerden lineer olmayan bir çekirdeğe hareket ettirerek yüksek boyutlu bir özellik uzayına eşlemektir.

Avantajlarını aşağıdaki gibidir;

- Hem doğrusal, hem de doğrusal olmayan verilere uygulanır.
- Yüksek doğruluk oranına sahiptir.
- Karmaşık hiper düzlemleri modelleyebilir.
- Çok sayıda bağımsız değişken ile çalışabilir.
- Aykırı gözlemlere karşı dayanıklıdır.

Dezavantajları ise;

- Olasılıksal tahminler üretmezler.

- Çekirdek (kernel) fonksiyonları pozitif tanımlı sürekli fonksiyonlar olmalıdır.
- Çekirdek (kernel) fonksiyonu eğitim ve tahminlemede kullanılır. Kullanılan dört farklı çekirdek fonksiyonu vardır; doğrusal (linear), polinom (polynomial), radyal temel (radial basis), sigmoid olarak bilinmektedir (Patle ve Chouhan, 2013).

Doğrusal: $\mathrm{K}(\mathrm{x}, \mathrm{x_j}) = \mathrm{x} \,.\, \mathrm{x}^{\mathrm{T}}$

Polinom: $\mathrm{K}(\mathrm{x}, \mathrm{x_j}) = (1 + \mathrm{x} \,.\, x_i^T)^d$

Radyal tabanlı: $\mathrm{K}(\mathrm{x}, \mathrm{x_j}) = \mathrm{e}^{-\gamma\|\mathrm{x}-\mathrm{x_j}\|^2}$

Sigmoid: $\mathrm{K}(\mathrm{x_j}, \mathrm{x_j}) = \tanh(\mathrm{jX}_i^T\mathrm{X_j} + \mathrm{r})$

3.3 Rassal Ormanlar Regresyonu (RF)

Rassal ormanlar, adından da anlaşılacağı gibi, bir topluluk olarak çalışan çok sayıda bireysel karar ağacından oluşur. Rassal ormanlardaki her bir ağaç bir sınıf tahmini verir ve en çok oyu alan sınıf, modelin öngörüsü haline gelir.

Torbalama(bagging) 1996 yılında Breiman tarafından ortaya konulmuştur. Diğer bir adı ile Bootstrap aggregation (bagging), istatistiksel bir öğrenme yönteminin varyansını azaltmaya yönelik bir yöntemdir. Burada her bir ağacın yetiştirileceği eğitim setindeki örneklerden rastgele bir seçim (değiştirilmeden) yapılır. Torbalama, sayısal bir sonuç tahmin ederken tahminlerin ortalamasını alır ve bir sınıf tahmin ederken oylamaya göre tahmin yapar (Breiman, 1996). Bir bootstrap örneği ile iadeli olarak eğitim setinden eşit oranda m tane örnek içeren örneklemler üretilir. T boostsrap örneklemleri B_1, B_2, … ,B_T üretilir ve her bir bootstrap örneği B_i için bir C_i, sınıflandırıcısı oluşturulur. Son sınıflandırıcı olan C^*, C_1, C_2, …C_T sınıflandırıcılarının en çok tahmin ettiği sınıfı baz alarak elde edilen sınıflandırıcıdır (Ekelik ve Altaş, 2019).

Rassal ormanların Bagging'den farkı gözlem ve değişken seçiminde rasgele seçimler yapmasından ortaya çıkmaktadır. RF, torbalama (bagging) ile rastgele alt uzay yöntemi (random subspace) yöntemlerinin birleşimi ile oluşmuştur. Bir topluluk öğrenme (ensemble learning) yöntemidir. Torbalama yönteminin gelişmiş bir şekli

olarak kabul edilebilir. Bu yöntemi Breiman makalesinde şöyle açıklamaktadır: k.'ıncı ağaç için, geçmiş rastgele vektörlerden $\theta_1, \ldots, \theta_{k-1}$ bağımsız ancak aynı dağılıma sahip θ_k rastgele vektörü oluşturulur ve her ağaç x girişindeki en popüler sınıf için bir birim oy kullanarak yani x'in giriş vektörü olduğu h(x, θ_k) sınıflandırıcısının sonucu olarak eğitim veri seti ve θ_k kullanılarak oluşturulur (Breiman, 2001). Örneğin, torbalamada, θ rastgele vektör, *N* eğitim veri setindeki gözlem sayısı olmak üzere, kutularda rastgele *N* gözlemlerinden elde edilen *N* kutudaki sayımlar olarak oluşturulur. Rastgele bölünme seçiminde θ, 1 ile *k* arasında bir dizi bağımsız rastgele tam sayıdan oluşur. θ 'nın yapısı ve boyutsallığı, ağaç yapımındaki kullanımına bağlıdır (Breiman, 2001). Aşağıdaki maddeler rassal ormanlar algoritmasının işleyişi hakkında fikir sunmaktadır (Hastie vd., 2001).

- Ağaçlar için gözlemler boostrap rastgele örnek seçim yöntemi ile değişkenler rastgele alt uzay yöntemi (random subspace) ile seçilir.
- Karar ağacının her bir düğümde en iyi dallara ayırıcı (bilgi kazancı) değişken tüm değişkenler arasından rastgele seçilen daha az sayıdaki değişken arasından seçilir.
- Ağaç oluşturmada veri setinin 2/3'ü kullanılır. Dışarıda kalan veri ağaçların performans değerlendirmesi ve değişken öneminin belirlenmesi için kullanılır.
- Her düğüm noktasında rastgele değişken seçimi yapılır (regresyonda p/3, sınıflamada √p).
- Nihai tahmin için ağaçlardan tahmin değerleri talep edilirken her bir ağacın daha önce hesaplanan hata oranları göz önüne alınarak ağaçlara ağırlık verilir.

Literatürde rassal ormanlarla ilgili çalışma yapan örnekler şunlardır: Dietterich 1998 yılında geliştirdiği her düğümde bölmenin K en iyi bölme arasından rastgele seçildiği rastgele bölme seçimidir. Breiman ise 1999 yılında, orijinal eğitim setindeki çıktıları rastgele hale getirerek yeni eğitim setleri oluşturur. Diğer bir yaklaşım, eğitim setindeki örnekler üzerindeki rastgele bir ağırlıklandırma sonucu eğitim setini seçmektir. Amit ve Geman 1997 yılında yazılı karakter tanıma üzerine önemli bir makalede çok sayıda geometrik özellik tanımlar ve her bir düğümdeki en iyi bölünme

için bunların rastgele bir seçimini araştırır. Ho 1998 yılında, her bir ağacı büyütmek için kullanılacak bir özellik alt kümesinin rastgele seçimini yapan "rastgele alt uzay" yöntemi hakkında birçok makale yazmıştır.

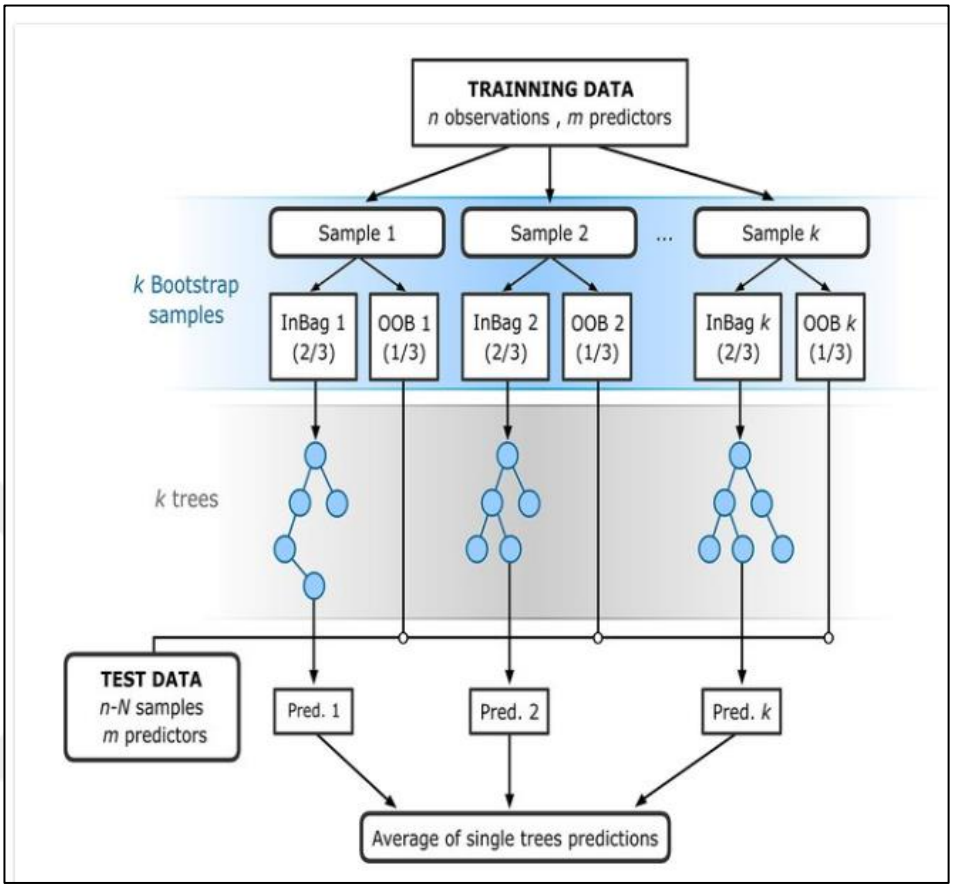
Breiman (2001), sınıflandırma ve regresyon ağaçları (CART) tekniğini kullanarak karar ağaçlarının inşası sırasında ek rastgelelik ortaya koymuştur. Bu tekniği kullanarak, her bir iç düğümde seçilen özniteliklerin alt kümesi, Gini indeks buluşsal yöntemleri ile değerlendirilir. En yüksek Gini indeksine sahip özellik, o düğümdeki bölünmüş özellik olarak seçilir. Gini indeksi Breiman, Friedman, Olshen ve Stone (1984) tarafından tanıtılmıştır. Ancak, ilk olarak 1912'de İtalyan istatistikçi Corrado Gini tarafından tanıtılmıştır. Endeks, verilerin saflığını ölçmek için kullanılan bir fonksiyondur, yani bir olayın gerçekleşip gerçekleşmeyeceğinden ne kadar belirsiz olduğunu ifade etmektedir. Sınıflandırmada bu olay sınıf etiketinin belirlenmesi olacaktır (Fawagreh, Gaber ve Elyan, 2014). Gini endeksi aşağıdaki gibi hesaplanmaktadır;

$$Gini(t) = 1 - \sum_{i=1}^{N} P(C_i|t)^2 \qquad (3.1)$$

t bir koşul, N veri kümesindeki sınıf sayısı ve C_i, veri kümesindeki i. sınıf etiketidir. RF ile ilgili orijinal makalede (Breiman, 2001), RF hata oranının korelasyon ve güce bağlı olduğu gösterilmiştir. RF'deki herhangi iki ağaç arasındaki korelasyonun arttırılması, orman hata oranını arttırır. Düşük hata oranına sahip bir ağaç güçlü bir sınıflandırıcıdır. Tek tek ağaçların gücünü artırmak, RF hata oranını azaltır (Fawagreh, Gaber ve Elyan, 2014).

Rassal ormanlarda çok sayıda ağaç olması durumundan dolayı, torba dışı (out of bag-OOB) hata payının olduğu gözlemlenmiştir. Oluşturulan her ağaca önceden hesaplanan OOB hata oranına göre bir ağırlık verilir. En düşük hata oranına sahip ağaç en yüksek ağırlığı, en yüksek hata oranına sahip ağaç en düşük ağırlığı alır. Her ağaç belirlenen ağırlığa göre yaptığı sınıf tahmini için bir oy verme işlemine tabi tutulur. Ağaçların oyu oluştuktan sonra rassla orman (random forest) algoritmasında bu

ağırlıklı oylar toplanır (Kıvrak, 2021). İşlem için bir akış şeması Şekil 3.5'te görülmektedir.



Şekil 3.5 Rassal ormanlar regresyonu akış diyagramı (Rodriguez-Galiano vd., 2016)

Son yıllarda neredeyse tüm disiplinlerde birçok RF uygulaması geliştirilmiştir. RF özet uygulama alanları Tablo 3.1'de görülmektedir.

Tablo 3.1 RF uygulama alanları (Fawagreh, Gaber ve Elyan, 2014)

Referanslar	Sektör adı	Amaç	Özet bilgi
Cutler vd. (2007)	Ekoloji	Tür sınıflandırması	Çeşitli türleri sınıflandırmak için RF doğruluğunu diğer tekniklerle karşılaştırmıştır. RF diğerlerinden üstün olduğunu kanıtlamıştır.
Klassen vd. (2008)	Sağlık	Kanser sınıflandırması	RF'nin kanseri kesin olarak sınıflandırabildiğini kanıtlamak için deneyler yapmıştır. Ayrıca, RF'nin birkaç farklı gen seti ile hız ve doğruluk açısından mikrodizi verileri için iyi performans gösterdiğini bulmuştur.
Hu (2009)	Sağlık	Meme kanserinde patolojik tam yanıtın tahmini	Meme kanserinde patolojik tam yanıtın tahminini incelemek için uygulanmıştır. Sonuçlar, RF'nin özellik seçim şemasının biyolojik önemi olan önemli genleri tanımlayabildiğini göstermiştir.
Gao vd. (2009)	Astronomi	Astronomik nesne sınıflandırması	Çok dalga boylu veri sınıflandırması üzerine bazı deneyler yapmıştır. Sonuçlar, RF'nin astronomik nesne sınıflandırması için etkili olduğunu kanıtlamıştır.
Flaxman vd. (2011)	Otopsi	Ölüm nedeni tahmini	Ölüm nedenini tahmin etmek için RF kullanan yeni bir CCVA (bilgisayar kodlu sözlü otopsi) yöntemi tanıtılmıştır.
Zaklouta vd. (2011)	Trafik ve ulaşım planlaması	Trafik işaretleri sınıflandırması	Farklı boyutta HOG tanımlayıcıları ve mesafe dönüşümleri kullanarak 43 tür trafik işaretini sınıflandırmak için K-d ağaçları ve RF'ler kullanılmıştır.
Löw vd. (2012)	Tarım	Ekinlerin doğru sınıflandırılması	Mahsul sınıflandırma doğruluğunu iyileştirmek ve harita belirsizliği hakkında uzamsal bilgi sağlamak için RF ve SVM sınıflandırıcılarının bir kombinasyonunu kullanılmıştır.
Boulesteix vd. (2012)	Biyoinformatik	Biyoinformatikte son gelişmeler araştırması	Bu alandaki RF uygulamalarının temsili örnekleri de dahil olmak üzere 10 yıllık RF geliştirmesi birleştirilmiştir.

3.4 Gradyan Arttırma Makineleri (GBM)

Orijinali sınıflandırma problemlerine odaklanan çok sayıda güçlendirme algoritması çeşidi vardır (Kuhn ve Johnson, 2013). 1990'lar boyunca, en başarılısı AdaBoost algoritması olmak üzere birçok yaklaşım geliştirilmiştir (Freund ve Schapire, 1999). 2000'de Friedman, AdaBoost'u önemli istatistiksel kavramlarla (örneğin, kayıp fonksiyonları ve toplamsal modelleme) ilişkilendirdi ve bu onun boosting çerçevesini regresyon problemlerine ve çoklu kayıp fonksiyonlarına genelleştirmesine izin vermiştir (Friedman, 2001). Gradyan artırma algoritması (GBM), ilk önce AdaBoost algoritmasının tanıtılmasıyla en kolay şekilde açıklanabilir. AdaBoost algoritması, her gözleme eşit ağırlık atanan bir karar ağacını eğiterek başlar. İlk ağacı değerlendirdikten sonra, sınıflandırılması zor olan gözlemlerin ağırlıklarını arttırmaktadır ve sınıflandırması kolay olanların ağırlıklarını azaltmaktadır. İkinci ağaç bu nedenle bu ağırlıklı veriler üzerinde büyütülür. Buradaki fikir, ilk ağacın tahminlerini geliştirmektir. Bu nedenle yeni modelimiz Ağaç 1 + Ağaç 2'dir. Daha sonra bu yeni iki ağaç topluluğu modelinden sınıflandırma hatası hesaplanır ve revize edilen artıkları tahmin etmek için üçüncü bir ağaç oluşturulur. Bu işlem belirli sayıda yinelenir. Sonraki ağaçlar, önceki ağaçlar tarafından iyi sınıflandırılmamış gözlemleri sınıflandırmamıza yardımcı olur. Bu nedenle, nihai topluluk modelinin tahminleri, önceki ağaç modelleri tarafından yapılan tahminlerin ağırlıklı toplamıdır (Singh, 2022). Birkaç denetimli makine öğrenimi algoritması, tek bir tahmine dayalı modele dayanır, örneğin; doğrusal regresyon, tek karar ağaçları ve destek vektör makineleri. Torbalama (bagging) ve rastgele ormanlar (random forest) ise, birden fazla modeli genel bir topluluk halinde birleştirerek çalışır. Topluluğu oluşturan bireysel temel modellerden gelen tahminleri birleştirerek (örneğin, regresyonda ortalama alarak) yeni tahminler yapılır. Ortalama alma varyansı azalttığı için, torbalama (ve dolayısıyla rastgele ormanlar) en etkili şekilde düşük sapmalı ve yüksek varyanslı modellere uygulanır. Güçlendirme, daha basit modellerden (tipik olarak karar ağaçları) bir topluluk oluşturmak için genel bir algoritma olsa da, yüksek yanlılık ve düşük varyanslı modellere daha etkili bir şekilde uygulanır. Hızlandırma, torbalama gibi herhangi bir modele uygulanabilmesine rağmen, genellikle en etkili şekilde karar ağaçlarına uygulanır (Hastie vd., 2001).

GBM, AdaBoost'un sınıflandırma ve regresyon problemlerine kolayca uyarlanabilen genelleştirilmiş versiyonudur. Gradyan arttırma tek bir tahminsel model formunda olan modeler serisi oluşturur. Seri içerisindeki bir model serideki bir önceki modelin tahmin artıklarının (hatalarının) üzerine kurularak oluşturulur. GBM diferansiyellenebilen herhangi bir kayıp fonksiyonunu optimize edebilen gradyan azaltma (gradient descent) algoritmasını kulllanmaktadır. Şekil 3.6'da gradyan arttırma modelinin algoritma adımları görülmektedir. Algoritma içerisindeki fonksiyonları tanımlayacak olursak; y_i tahmin etmek istenilen sütun, x_i ise y_i gözlemlerini tahmin etmek için kullanılan niteliklerdir ve aynı zamanda N örnek sayısını, L ise kayıp fonksiyonu ifade etmektedir. Birinci adım olarak belirtilen ifadede, her satır için modelin tahmin artıklarını toplayarak *arg min* ile bu toplamı minimuma döndürecek tahmin değerlerini bulmak gerektiğini anlatmaktadır. Bunun için türev alınmaktadır bu da direkt olarak tahmin edilenden gerçekleşen değerleri çıkararak sıfır'a eşitleyip tahmin değerini bulmaya işaret etmektedir. Yukarıda da bahsedildiği üzere kayıp fonksiyonunu minimize etmek için gradyan azaltma kullanılır. Bu işlemler kayıp fonksiyonu ile yapılan ilk tahmin değerlerinin gerçek değerlerin ortalaması olması gerektiğini kanıtlamaktadır. İkinci adımda ise, görülen M değeri ağaç sayısını ifade etmektedir. Her bir örnek için kayıp fonksiyonunun x'e göre türevi alınırsa birinci adımda bahsedildiği gibi tahmin değerinden gerçek değerin çıkarılmış haline eşit olduğu görülmektedir ve bu değeri -1 ile çarpımı gerçek değerin tahmin değerinden çıkarılmasına eşit olmaktadır, yani hataların sütunua denk gelmektedir. $f = f(m-1)$, tahmin fonksiyonunda bir önceki ağacın tahmin değerinin uygulanmasıdır yani, $Gerçek\ değer - F(m-1)(x))$, birinci ağaç için önceki adımda kullanılan ortalamanın içine yazılmıştır. *For* döngüsünü örnekler üzerinden döndürmek her örnek için tek tek hata hesaplanması anlamına gelmektedir. (b) bölümünde ise tahminler yerine hatalar üzerinden regresyon ağacı kurulması gerekmektedir. Burada terminal bölgeler olarak bahsedilen ağacın bittiği yerler, yani tahmin edilen hataların yazıldığı yapraklar *j* olarak ifade edilmektedir. (c) kısmında, ilk adımda yapılan minimizasyonun jenerik hali ve artık önceki tahmin hesaba katılmaktadır. Burada *R(j,m)*'deki x_i yapraktaki değer olarak görülmektedir. Bu bölümde ortalama alabilmenin kolaylığı yine kullanılan kayıp fonksiyonundan gelmektedir. (d) bölümünde, her örnek için yeni tahmin yürütülmesi gerekmektedir.

f(m-1) fonksiyonu bir önceki ağaçta aynı örneğin tahmini, ikinci kısım da o örneğin bulunduğu yapraktaki hatayı ifade etmektedir. Bu adımı belirlenen ağaç sayısı kadar tekrarladıktan sonra üçüncü adıma geçilmektedir. Üçüncü adımda, her örnek için yapılan tahminin son ağaçta bulunan tahmin olduğunu belirtmektedir (Hastie vd., 2001).

1. Initialize $f_0(x) = \arg\min_\gamma \sum_{i=1}^{N} L(y_i, \gamma)$.
2. For $m = 1$ to M:
   (a) For $i = 1, 2, \ldots, N$ compute
   $$r_{im} = -\left[\frac{\partial L(y_i, f(x_i))}{\partial f(x_i)}\right]_{f=f_{m-1}}.$$
   (b) Fit a regression tree to the targets r_{im} giving terminal regions R_{jm}, $j = 1, 2, \ldots, J_m$.
   (c) For $j = 1, 2, \ldots, J_m$ compute
   $$\gamma_{jm} = \arg\min_\gamma \sum_{x_i \in R_{jm}} L\left(y_i, f_{m-1}(x_i) + \gamma\right).$$
   (d) Update $f_m(x) = f_{m-1}(x) + \sum_{j=1}^{J_m} \gamma_{jm} I(x \in R_{jm})$.
3. Output $\hat{f}(x) = f_M(x)$.

Şekil 3.6 Gradyan arttırma (GBM) algoritma adımları (Hastie vd., 2001)

3.4.1 Hiperparametreler

Basit bir GBM modeli, iki hiperparametre kategorisi içerir: artırma hiperparametreleri ve ağaca özgü hiperparametreler. İki ana güçlendirme hiperparametresi şunları içerir:

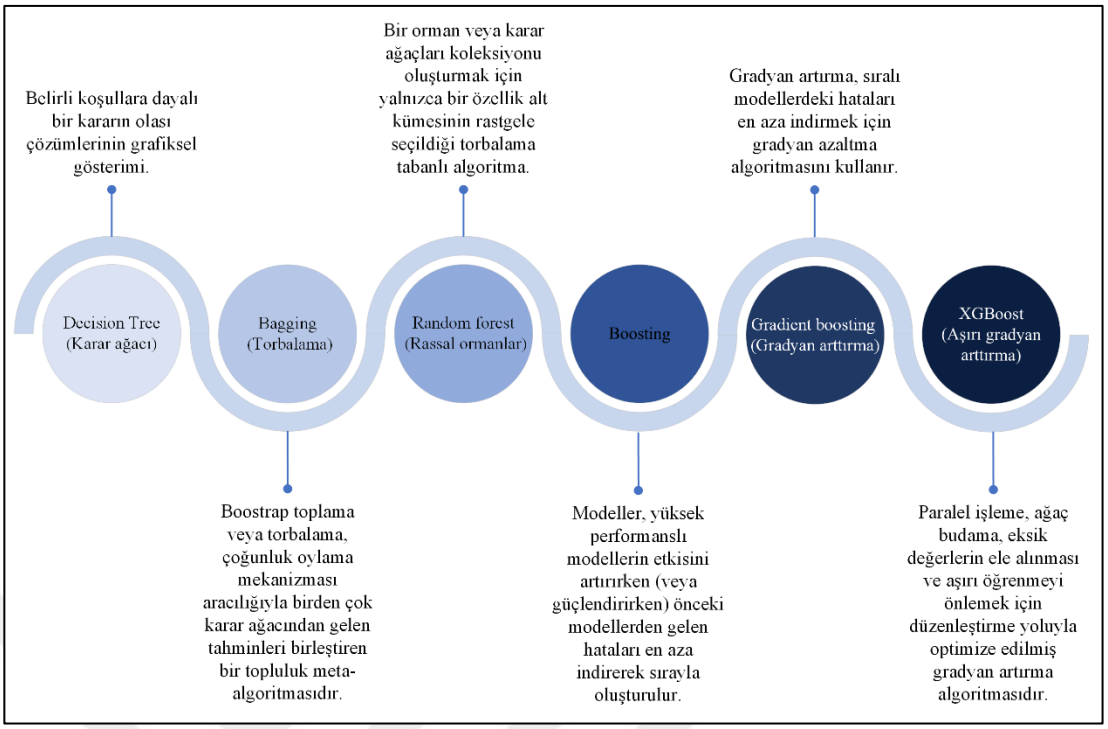
- ***Ağaçların sayısı (number of trees):*** Torbalama ve rastgele ormanlarda bağımsız olarak yetiştirilen ağaçların ortalamasıdır. Bununla birlikte, geçmiş ağacın hatalarını düzeltmek için her ağaç sırayla büyüdüğünden GBM'ler farklı şekilde çalışır. Örneğin, regresyonda GBM'ler kullanıcı izin verdiği sürece artıkları takip edecektir. Ayrıca, diğer hiperparametrelerin değerlerine bağlı olarak GBM'ler

genellikle birçok ağaç gerektirir. İlgili kayıp fonksiyonunu en aza indiren optimum ağaç sayısını çapraz doğrulama ile bulunmalıdır.

- ***Öğrenme oranı (learning rate):*** Her ağacın nihai sonuca katkısını belirler ve algoritmanın gradyan azaltmada ne kadar hızlı ilerlediğini öğrenir. Değerler 0–1 arasında ve tipik değerler 0,001–0,3 arasındadır. Daha küçük değerler, modeli her bir ağacın kendine özgü özelliklerine karşı sağlam kılar ve böylece iyi bir genelleme yapmasına izin verir. Daha küçük değerler, aşırı uymadan (overfitting) önce durmayı da kolaylaştırır; ancak, sabit sayıda ağaçla optimuma ulaşamama riskini artırırlar ve hesaplama açısından daha zahmetlidirler. Bu hiperparametreye büzülme (shrinkage) de denir. Genel olarak, bu değer ne kadar küçük olursa, model o kadar doğru olabilir, ancak aynı zamanda dizide daha fazla ağaç gerektirecektir (Greenwell ve Boehmke, 2019).

3.5 Aşırı Gradyan Arttırma Makineleri (XGBoost)

XGBoost, GBM'in hız ve tahmin performansını arttırmak üzere optimize edilmiş; ölçeklenebilir ve farklı platformlara entegre edilebilir halidir. Aşırı öğrenmeye karşı duyarlı, genellikle sınıflandırma ve regresyon uygulamalarında kullanılan ve son zamanlarda makine öğrenmesi alanında adından çokça bahsedilen bir algoritma olarak bilinir. Karar ağacı temelli algoritmalar genellikle sınıflandırma ve regresyon problemlerinde tercih edilebilir. Bu algoritmalar zaman içerisinde birçok evrime uğramıştır. Bu evrimi Şekil 3.7'deki zamana akışında görülmektedir.



Şekil 3.7 Karar ağacı algoritmalarının XGBoost'a evrimi

XGBoost, ilk olarak Washington Üniversitesi'nde iki araştırmacı olan Tianqi Chen ve Carlos Guestrin tarafından SIGKDD (Uluslararası Bilgi Keşfi ve Veri Madenciliği Konferansı) 2016 konferansında sunulmuştur. Sadece akademik anlamda değil aynı zamanda Google LLC'nin bir yan kuruluşu olan, çevrimiçi bir veri bilimcileri ve makine öğrenimi uygulayıcıları topluluğu olan Kaggle yarışmalarının da yıldızı haline gelmiştir. XGBoost ve GBM topluluk temelli zayıf öğrenenler kullanarak zayıf öğrenenler gradient-descent (gradyan azaltma) metodu ile desteklenmektedir. Ancak XGBoost, sistem optimizasyonu ve algoritmik geliştirmeler yoluyla temel GBM çerçevesini geliştirmektedir (Morde, 2019).

3.5.1 Sistem Optimizasyonları

- ***Paralel çalıştırma:*** XGBoost, karar ağaçlarını oluşturuken paralelizasyon yaparak çok daha hızlı oluşturulmasını sağlamaktadır. Bunu yapabilmesinin altında temel-öğrenenleri (base-learners) oluştururken, iç ve dış döngüler arasında geçiş yapabiliyor olmasıdır. Normal şartlarda dış döngüler, karar ağacının yapraklarını oluştururken iç döngüler tamamlanamayacağı için yani öznitelikler

hesaplanmadan ağacın yaprakları oluşmayacağı için paralelizasyon sınırlanır. Bu nedenle, çalışma süresini iyileştirmek için döngülerin sırası, tüm örneklerin genel taraması yoluyla başlatma ve paralelizasyon ile sıralama kullanılarak değiştirilir. Bu anahtar, hesaplamadaki herhangi bir paralelleştirme ek yükünü dengeleyerek algoritmik performansı iyileştirir.

- ***Ağaç budama:*** GBM çerçevesinde ağaç dallarını ayırırken durdurma kriteri, bölünme noktasındaki negative kayıp kriterine bağlıdır. XGBoost ise önce "max_depth" parametresi ile ağacın derinliğini belirleyerek eğer ağaç aşağı yönde fazla ilerlediyse geriye doğru budama yapmaktadır. XGBoost derinliğe öncelik verdiği için karmaşıklığı, dolayısıyla hesaplama performansını önemli ölçüde arttırmaktadır.
- ***Donanım optimizasyonu:*** Bu algoritma, donanım kaynaklarını verimli kullanmak için tasarlanmıştır. Her iş parçacığı kendi içindeki dahili belleklerinde gradyan istatistiklerini depolamaktadır. Bu sayede, ön belleğin doluluğu göz önünde bulundurulmaktadır. Çekirdek dışı (out-of-core) bilgi işlem gibi diğer geliştirmeler, belleğe sığmayan büyük veri çerçevelerini işlerken kullanılabilir disk alanını optimize eder.

BÖLÜM 4
RÜZGAR GÜCÜNÜN TAHMİNLENMESİ

Rüzgar enerjisi güç tahmini çalışmasında Türkiye'nin İç Anadolu bölgesinde bulunan bir rüzgar santralinin 2019 ve 2020 yıllarına ait 10 dakikalık verilerinden yararlanılmıştır. Veriler SCADA aracılığı ile temin edilmiştir. Rüzgar hızı ve rüzgar yönü değişkenleri kullanılarak güç tahmini yapılmıştır. Performansa göre tahmin modeline karar verilmiştir. Bu çalışmada 70 adet rüzgar türbinine sahip karasal rüzgar santrali verileri kullanılmıştır ve ayrıca santral 2013 Aralık ayından beri faaliyet göstermektedir. Bu santralde 60 adet E-82 E2 ve 10 adet E-82 E3 rüzgar türbini bulunmaktadır. Türbin tipleri farklı gibi görünse de fiziksel olarak homojendir. Bu çalışmada iki sene boyunca 10 dakikalık kayıt alan 1 adet E-82 E3 türbin tipine sahip rüzgar türbinin verilerinden faydalanılmıştır. Bu seçimi yaparken dikkate alınan kriterler arasında iki yıl boyunca bütün türbinlerin güç performansı incelenmiş ve büyük komponent değişimine uğramamış bakım performansı dikkate alınarak seçim gerçekleşmiştir. Veri gizliliği çerçevesinde sadece SCADA verilerine erişim izni alındığı için bakım performansı bilgisi paylaşılmamıştır. Yapılan tüm analizlerde R 4.0 proglamlama dili kullanılmıştır.

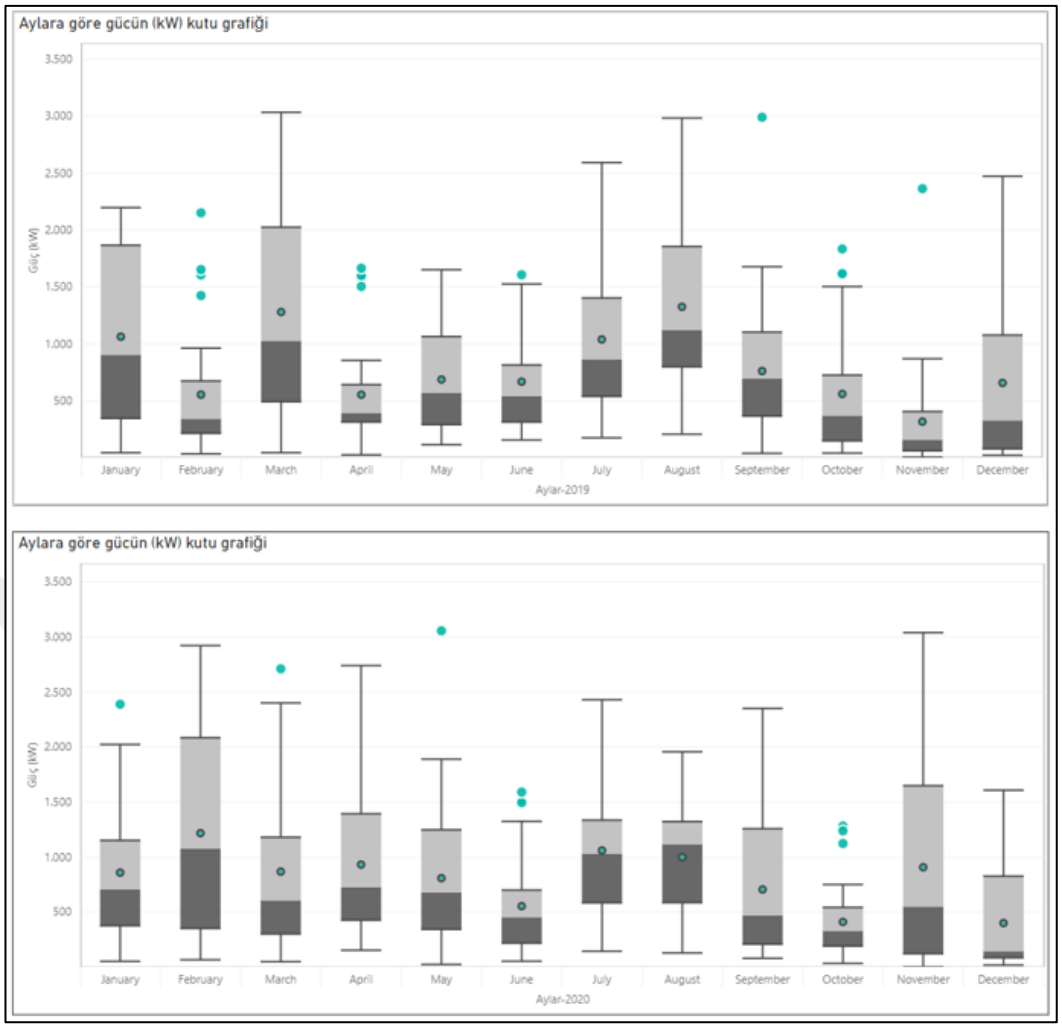
4.1 Veri Seti ve Analizi

SCADA üzerinden elde edilen verilerde rüzgar hızı, rüzgar yönü ve güç bilgisi yer almaktadır. Toplam veri seti 89.596 gözlem içermektedir. Aykırı (anomali) değerleri veya eksik verileri çıkardıktan sonra, makine öğrenimi modellerini uygulamak için veri kümesinin geri kalanı, yani 88.847 veri noktası dikkate alınmıştır. Veri seti, tüm veri setinin %80'ini içeren eğitim seti ve veri setinin %20'sini içeren test seti olmak üzere iki bölümden oluşmaktadır. Güç tahmini için elimizde bulunan verilere ait bazı tanımlayıcı istatiksel bilgiler Tablo 4.1'de görülmektedir.

Tablo 4.1 Veri setinin tanımlayıcı istatistikleri

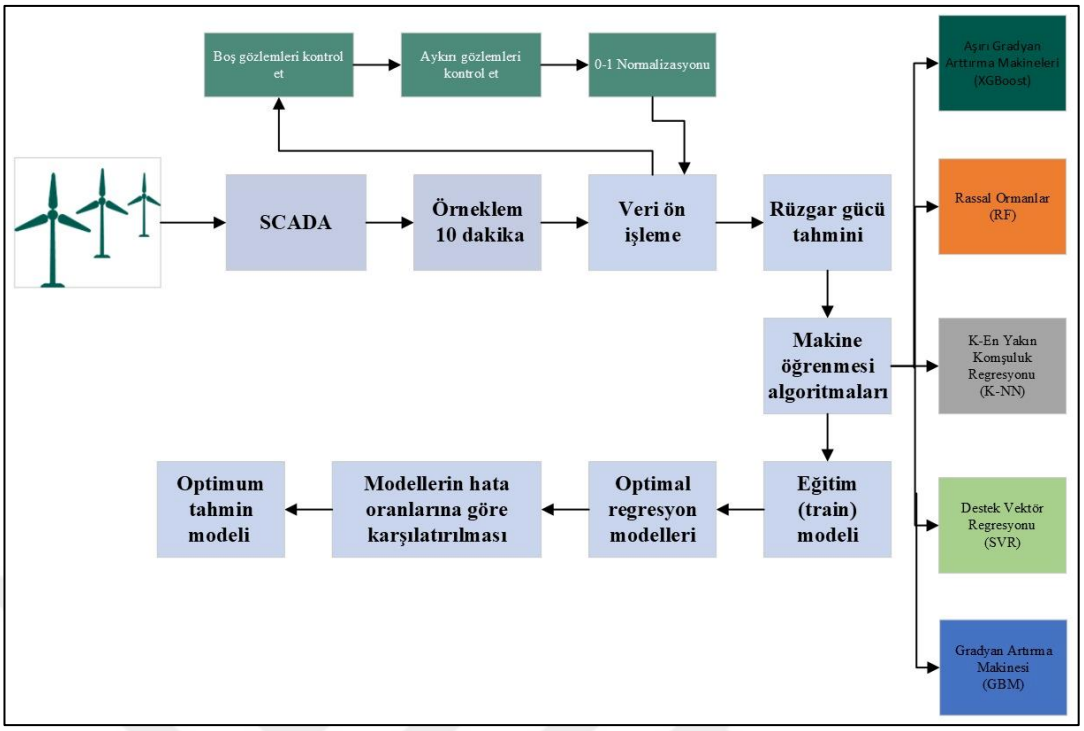
Tanımlayıcı istatistikler	**Rüzgar hızı (m/s)**	**Rüzgar yönü (º)**	**Güç (kW)**
Minimum	1,1	0	1
Birinci çeyreklik (p25)	4,80	46	130
Ortalama	7,84	132,70	858,50
Medyan	7,20	90	512
Üçüncü çeyreklik (p75)	10,50	218	1434
Maksimum (p100)	29,7	359	3071
Standart sapma	3,86	95,87	875,31

Şekil 4.1'de görüldüğü gibi 2019 ve 2020 yıllarına ait ay ölçütüne göre güç değişkeninin kutu grafikleri verilmektedir. Çeyrekler arası aralığın <1,5IQR (interquartile range) olan bıyık tipine göre kutu grafiği çizilmiştir. 2019 yılının Kasım ayının dağılım merkezi diğer ayların dağılımlarının en küçüğüdür, yani medyanı 159,78 kW olarak gözlenmiştir. 2020 yılında ise Aralık ayının dağılım merkezi en küçüktür (medyan 145,20 kW). Şekil 4.1'e göre, 2019 ve 2020 yıllarında en yüksek üretim ortalaması Ağustos ayına aittir. 2019 yılında Temmuz ve Eylül ayları, 2020 yılında ise Ocak ve Şubat aylarında simetrik dağılıma yakındır, hemen hemen her iki yarım kutu da aynı uzunluktadır. Bazı aylar potansiyel aykırı değerler içermektedir.



Şekil 4.1 Aylara göre gücün kutu grafiği

Şekil 4.2'de görüldüğü gibi, bu çalışmada kullanılan uygulamanın özet diyagramı gösterilmektedir. SCADA üzerinden verinin işlenmesinden kullanılan algoritmalara kadar bütün adımlar özetlenmiştir.

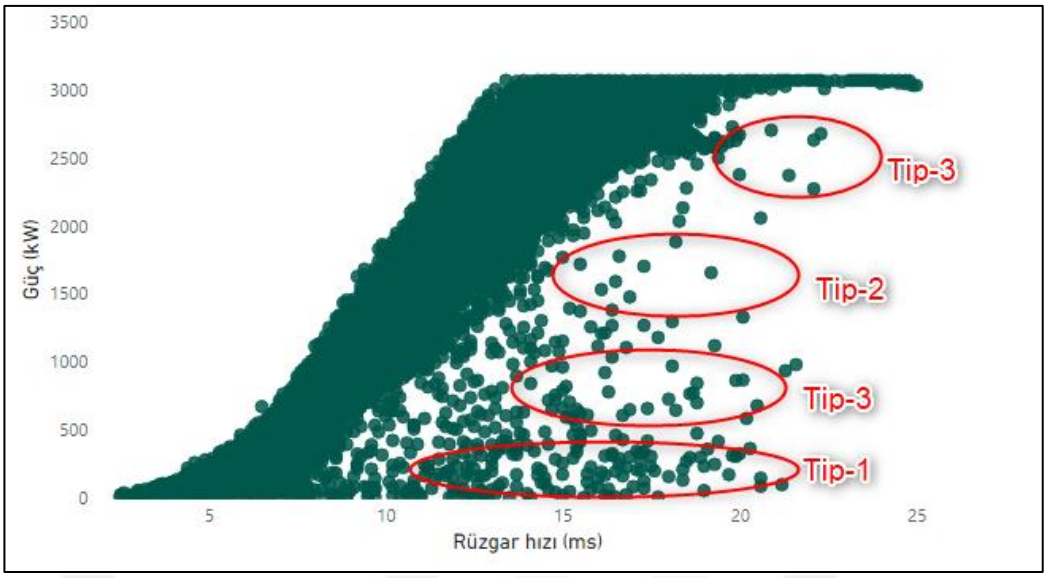


Şekil 4.2 Önerilen model için fonksiyonel diyagram

Veri analizi ve tahmini için izlenen adımlar Şekil 4.3'de gösterilmektedir. Veri keşfi, verilerin analizindeki ilk adımdır ve kullanıcıların dikkat noktalarını ve dikkate değer özellikleri keşfetmek için yapılandırılmamış bir şekilde büyük bir veri kümesini keşfettiği yerdir. Veri temizleme, verilerdeki alakasız, hatalı veya eksik kısımların tespit edilmesi ve ardından gereksinimlere uygun olarak verilerin değiştirilmesi ve çıkarılmasını ifade eder. Modelleme, tahmin etmek için makine öğrenimi algoritmasının eğitilmesini ve ardından en iyi modeli belirlemeye yardımcı olur. Makine öğrenimi algoritmasının performansı, eğitim ve test veri kümeleri kullanılarak farklı performans ölçütleriyle karşılaştırılır.

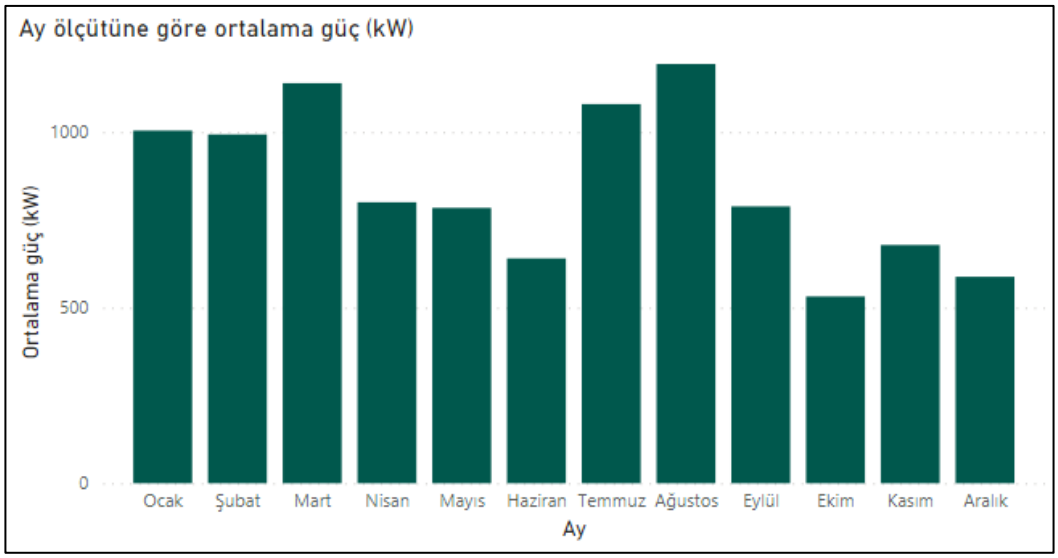


Şekil 4.3 Veri analizi ve tahmin yaparken gerçekleşen adımlar



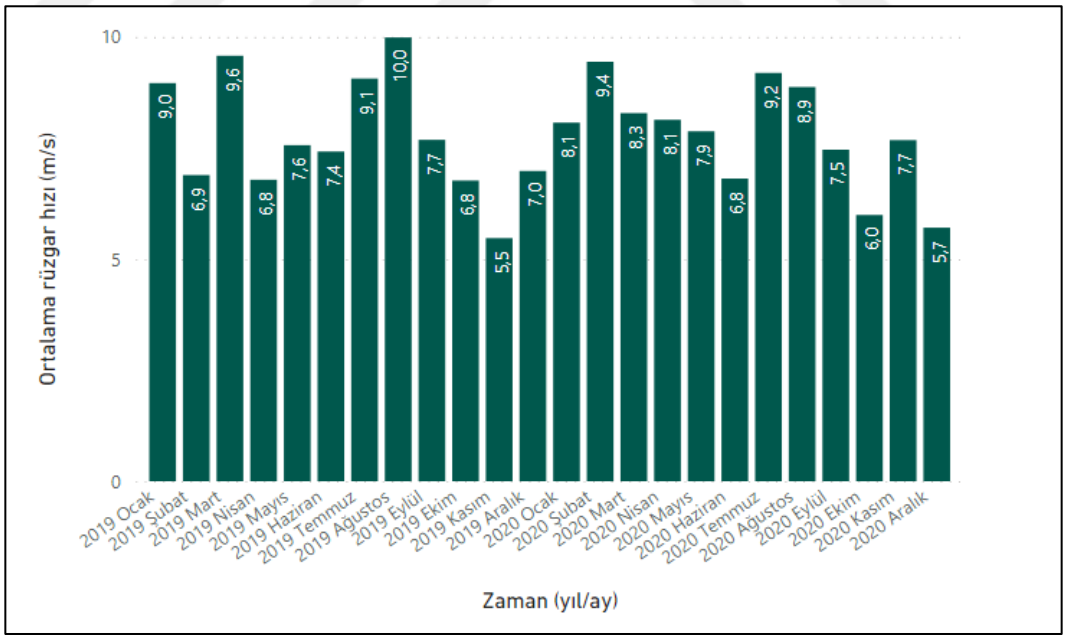
Şekil 4.4 Rüzgar hızı ve güç arasındaki ilişki-güç eğrisi

Rüzgar türbini gücünü daha yakından inceleyerek Şekil 4.4'te güç eğrisindeki bazı anomalilikler yorumlanmıştır. Tip-1 anomaliler, başlatma rüzgar hızından daha yüksek bir rüzgar hızında güç üretiminin sıfır olduğu yatay yoğun bir veri kümesi aracılığıyla dağılım grafiğinde görüntülenir. Bu tür anomaliler genellikle, bir çalışma günlüğü kullanılırken çapraz referans alınabilen türbin arıza süresi nedeniyle oluşur. Tip-2 anomaliler, rüzgar türbininin ideal güç eğrisinin altına düşen yoğun bir veri kümesiyle gösterilir. Bu anomaliler, türbin çıkış gücünün işletmeci tarafından işletme kapasitesinden daha düşük olacak şekilde kontrol edildiği güç kısıntısı nedeniyle ortaya çıkabilir. Güç kısıtlaması, büyük rüzgar enerjisi kapasitelerinin depolanmasındaki zorluk gibi çeşitli nedenlerle bir rüzgar santrali işletmecileri tarafından gerçekleştirilebilir. Tip-3 anomalileri, eğri etrafında keyfi olarak dağılır ve bunlar genellikle sensör bozulması veya arızasının bir sonucudur veya sinyal alma sırasındaki gürültüden kaynaklanabilir. Ayrıca, türbülansın bir sonucu olarak alınan tutarsız rüzgar hızı ölçümleri nedeniyle üretilen dağılım ile tip-2 ve tip-3 anomalilerinin bir bölümünün de gösterilebileceğini belirtmekte fayda vardır.



Şekil 4.5 Aylara göre ortalama gerçekleşen güç (kW)

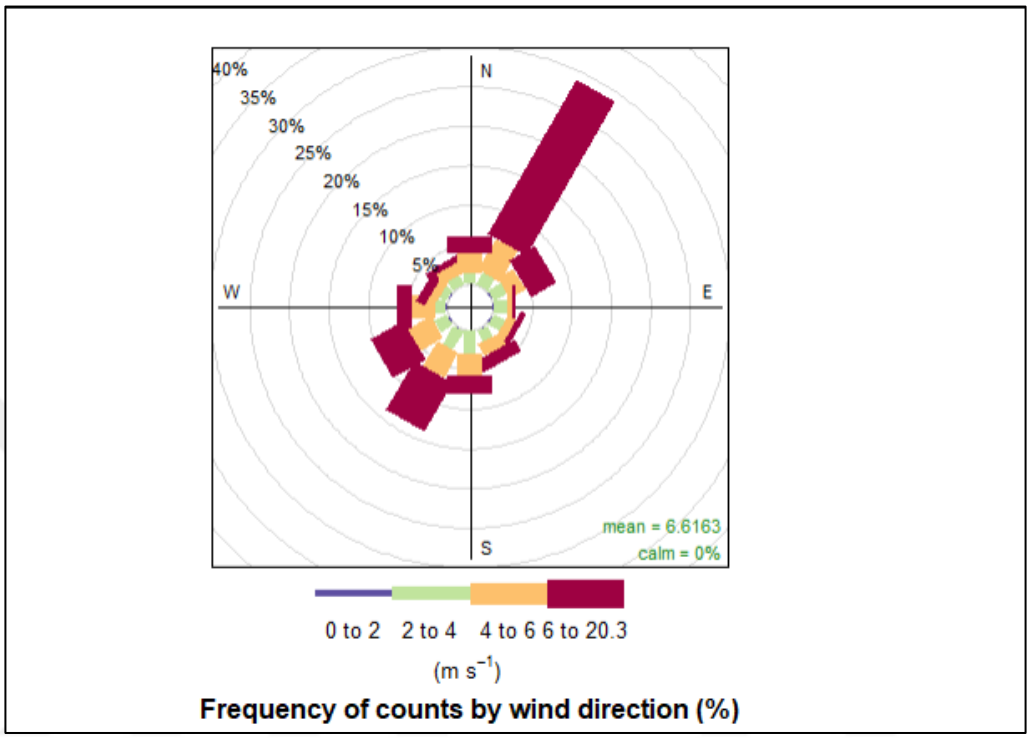
Şekil 4.5'te 2019 ve 2020 yıllarına ait aylara göre gerçekleşen güç verilerinde en fazla üretime sahip olan aylar sırasıyla; Ağustos, Mart ve Temmuz olarak kaydedilmiştir. Şekil 4.6'da görüldüğü üzere ortalama rüzgar hızları da aynı aylarda yani Ağustos, Mart ve Temmuz aylarında yüksek ortalamalara sahiptir.



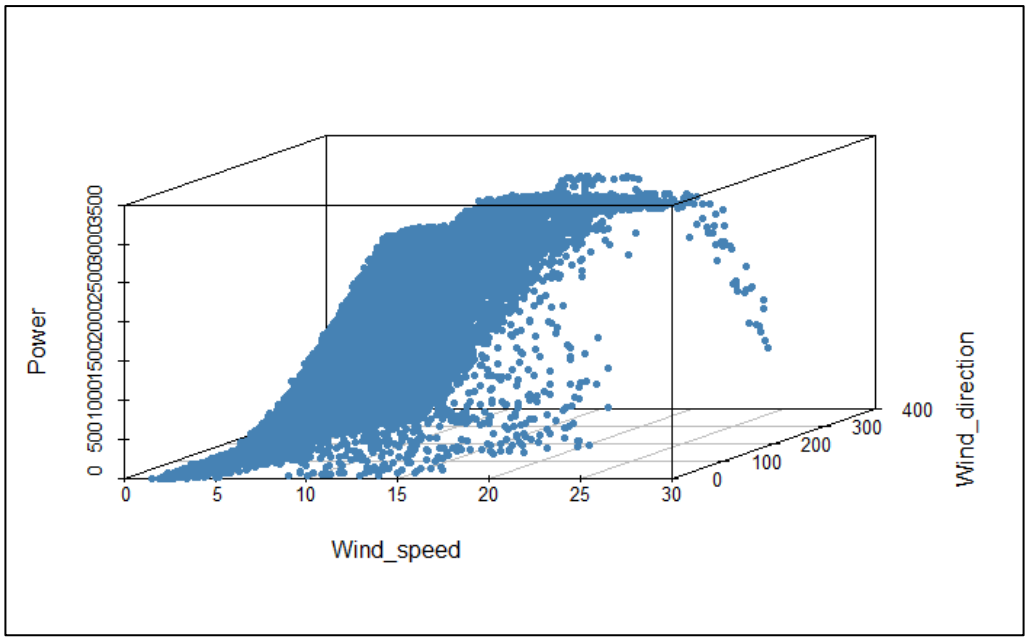
Şekil 4.6 SCADA 2019 ve 2020 verisi ile aylara göre ortalama rüzgar hızları (m/s)

Şekil 4.7'de görüldüğü gibi türbinin hakim rüzgar yönü kuzeydoğu 30º olarak gözlemlenmiştir. En fazla kuzeydoğu yönünden rüzgar esmektedir. Bu yönde esen

rüzgar yoğunluğunun maksimum güç üretimine katkı sağladığı Şekil 4.8'de görülmektedir. Rüzgar yönü 0º ile 30º arasında değiştiğinde güç üretimi de bu bölgelerde yoğun bir şekilde gözlemlenmektedir.

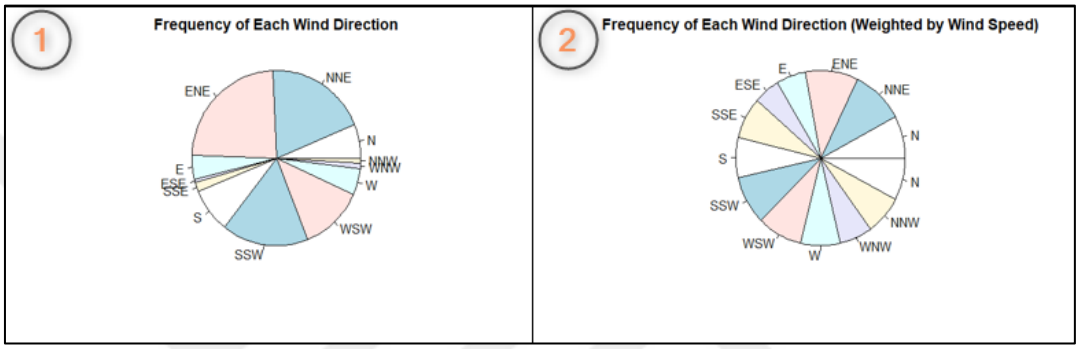


Şekil 4.7 Türbinin hakim rüzgar yönü



Şekil 4.8 Güç (kW), rüzgar hızı ve rüzgar yönü arasındaki üç boyutlu (3B) görsel

Rüzgar yönü kayıtlarındaki fark oldukça büyük gibi görünebilir, ancak Şekil 4.9'daki birinci grafikte sadece her 10 dakikalık aralıklarla rüzgar yönü dikkate alınmıştır ve rüzgar hızının etkisini göz ardı edilmiştir. Rüzgar bazı yönlerde daha sık esebilir, ancak diğer yönlerde daha hızlı esebilir. Doğal olarak, hız da dikkate alınmalıdır. Bundan dolayı, rüzgar hızına ağırlık vermek için her bir rüzgar yönünün frekansını her bir yöndeki ortalama rüzgar hızıyla çarparak ikinci grafik elde edilmektedir.

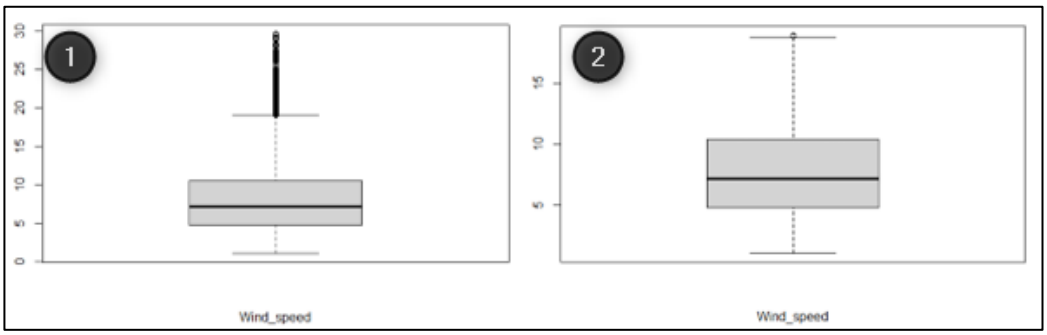


Şekil 4.9 Rüzgar hızı ile ağırlıklandırılmış rüzgar yönü frekansları

4.2 Veri Ön İşleme Prosedürleri

- ***Aykırı gözlemler:*** Bu çalışmada ilk olarak rüzgar santralinden elde edilen veri setinin kullanılabilirliği tüm türbinler için ayrı ayrı incelenmiştir. İncelenen veriler sonucunda arıza ya da bakım nedeniyle duruşa geçmiş bazı türbinlerden veri akışı sağlanamadığı görülmüştür ve en çok operasyonda kalan veriye sahip olan türbin, modellerde kullanılmak üzere seçilmiştir. Seçilen türbin üzerinden veri modelleri oluşturulmuş ve test edilmek üzere analiz edilmiştir. Seçilen türbin özelinde aykırı gözlemleri tespit etmek üzere R programlama dili aracılığıyla analizler gerçekleştirilmiştir. Kutu grafiği yöntemi ile aykırı gözlemler veri setinden çıkarılmıştır. Birinci ve üçüncü çeyreklikler belirlendikten sonra alt limit ve üst limitler belirlenmiş olup bu aralıklar dışında kalanlar ise aykırı gözlem olarak belirlenmiştir. Şekil 4.10'da görüldüğü gibi rüzgar hızına ait kutu grafiğinin aykırı gözlemlerini tespiti 1 numaralı grafikte verilmiştir. Ardından kutu grafiği yöntemi ile bu gözlemleri veri setinden çıkardıktan sonra tekrar kalan gözlemlerle kutu

grafiği çizdirilerek 2 numaralı grafik elde edilmiştir. Rüzgar hızının grafiğinden anlaşıldığı gibi aykırı gözlemler kapatma rüzgar hızının (25 m/s) üzerinde ve başlatma rüzgar hızının (3 m/s) altında seyretmektedir.

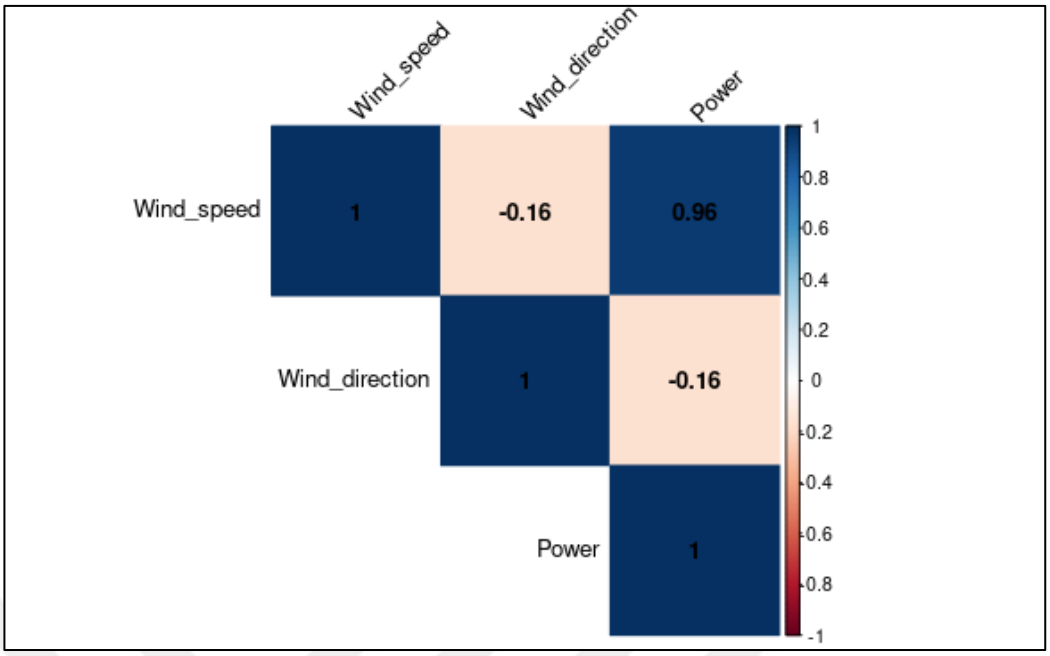


Şekil 4.10 Rüzgar hızının aykırı gözlem tespiti

- ***Normalizasyon (0-1):*** Rüzgar gücü tahmin modelinin girdi parametreleri, rüzgar hızını ve rüzgar yönünü içerir, ancak boyutları aynı büyüklük sırasına sahip değildir. Bu nedenle, bu girdi vektörlerinin aynı büyüklük sırasına göre düzenlenmesi esastır. Bu nedenle, girdi vektörlerini aşağıdaki gibi normalleştirmek için bir min-maks yaklaşımı kullanılmıştır:

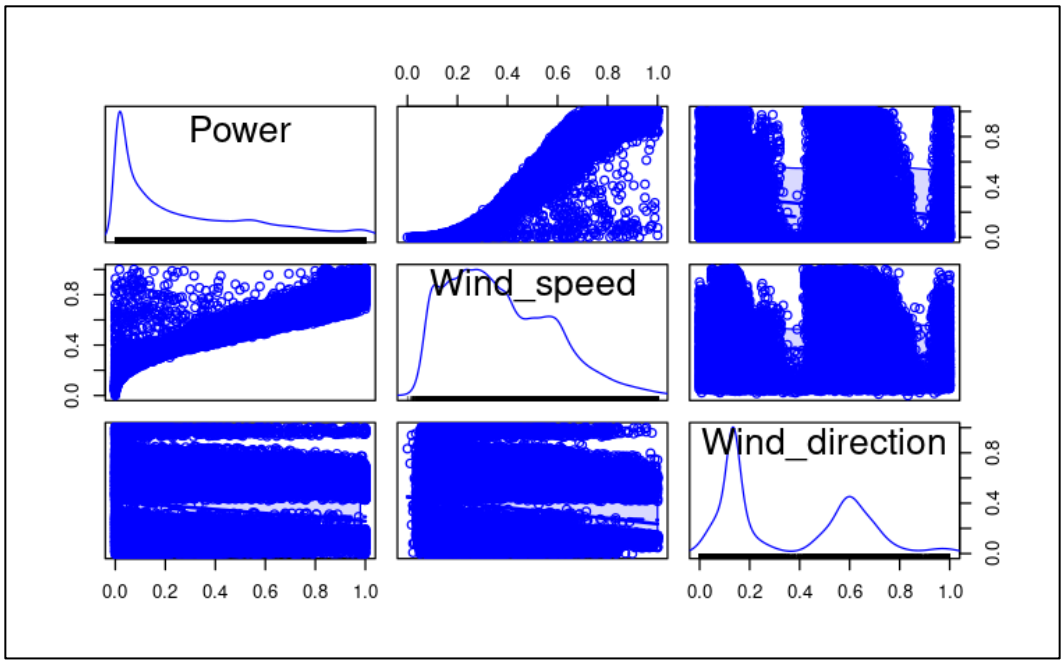
$$\bar{x} = \frac{x - x_{min}}{x_{max} - x_{min}} \tag{4.1}$$

Normalizasyon işlemlerinden sonra girdi (rüzgar hızı ve rüzgar yönü) ve çıktı (güç) değişkenleri arasındaki ilişki matrisleri Şekil 4.10 ve Şekil 4.11'de görülmektedir. Bilindiği üzere rüzgar hızı ile güç arasında güçlü bir ilişki vardır, eğitim veri setindeki bu değer %96 olarak görülmektedir.



Şekil 4.11 Girdi ve çıktı parametreleri arasında ilişki matrisi

Şekil 4.11 her bir değişkenin birbirleriyle olan ilişkisini açıklayan eşleştirilmiş dağılım grafiklerini göstermektedir. Şekil 4.12'de çapraz şekilli grafikler, her bir parametre özelliğinin olasılık dağılımını gösteren histogramları temsil eder.



Şekil 4.12 Serpme diyagramı (scatter plot) ile girdi ve çıktı parametreleri arasındaki ilişki

4.3 Sonuçların Değerlendirilmesinde Kullanılan Hata Fonksiyonları

Bu çalışmada, rüzgar enerjisi üretimini tahmin etmek için modellerin performanslarını değerlendirmek üzere ortalama mutlak hata (MAE), ortalama karesel hata (RMSE) ve belirtme katsayısı (R^2) kullanılmıştır. MAE, sürekli değişkenler için kullanılan bir model değerlendirme metriğidir. Tahmin hatası, bir örnek için gerçek değer ile tahmin edilen değer arasındaki farkla hesaplanır (Delgado ve Fahim, 2021). Kullanılan tüm bu ölçümlerin formülleri aşağıdaki gibidir:

$$MAE(y, \hat{y}) = \frac{1}{n}\sum_{i=0}^{n-1}|y_i - \hat{y}_i| \tag{4.2}$$

$$RMSE(y, \hat{y}) = \sqrt{\frac{1}{n}\sum_{i=0}^{n-1}(y_i - \hat{y}_i)^2} \tag{4.3}$$

$$R^2 = 1 - \frac{\sum_{i=0}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=0}^{n}(y_i - \bar{y})} \tag{4.4}$$

Performans ölçümlerinde y gözlemlenen değeri, $\hat{y}$ tahmin edilen değeri ve n toplam örnek sayısını temsil eder. MAE ve RMSE için, daha düşük tahmin değeri yüksek doğruluğa karşılık gelir ve 0 ile 1 arasındaki yüksek R^2 değeri nihai tahmin modelini belirlemek için kullanılmaktadır.

4.4 Rassal Ormanlar (RF) Regresyonu için Bulgular

Rassal ormanlar regresyonu için öncelikle default (kabul edilmiş) değerlerinden elde edilen çıktılarla modeli ayarlayarak elde edilmiş çıktılar kıyaslanarak en iyi rassal ormanlara ait parametrelere karar verilmiştir. Bunlar, mtry (her bölünmede kullanılan rasgele değişken sayısı), ntree (kullanılacak ağaç sayısı-boostrap örneklem sayısı) şeklindedir. Aşağıdaki koda göre, seçilecek olan değişken sayısını train seti içerisindeki değişken sayısının 3'e bölerek yaklaşık olarak bir değer elde ettikten

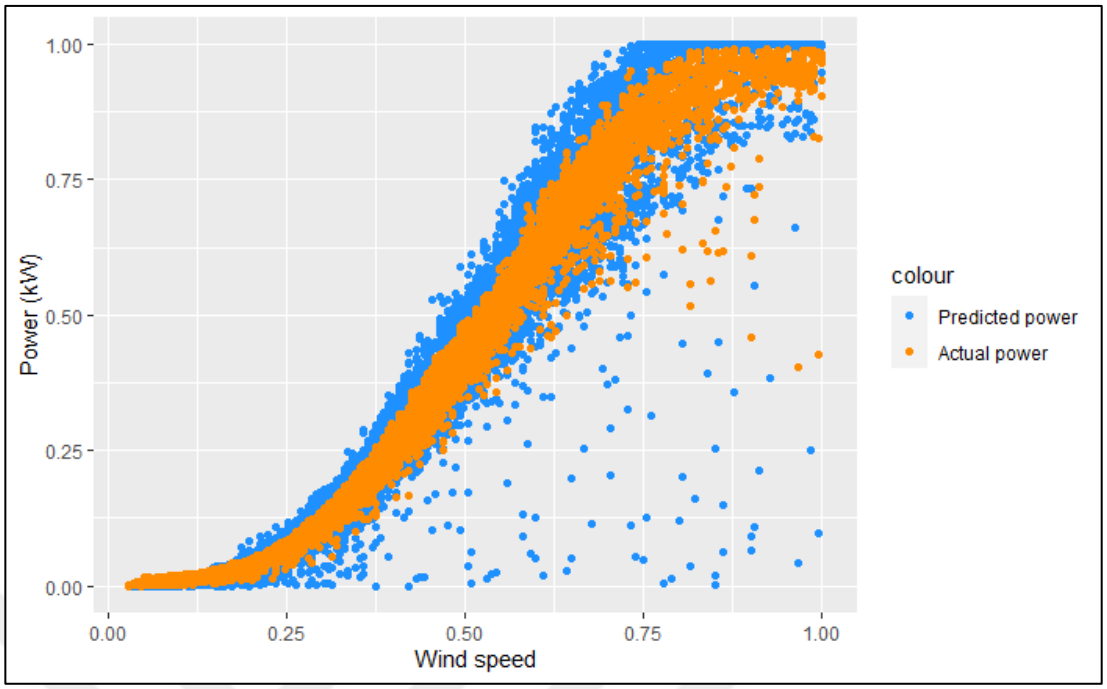
sonra, seçilecek değişken sayısını 1 ve 2 olarak belirleyerek denemeler yapması sağlanmıştır. Daha sonra optimal seçilecek olan değişken sayısı bulunmaktadır.

```
ctrl <- trainControl(method = "cv", number = 10)
ncol(train_x)/3
tune_grid <- expand.grid(mtry = c(1,2))
rf_tune <- train( train_x, train_y,
                  method = "rf",
                  tuneGrid = tune_grid,
                  trControl = ctrl)
rf_tune
plot(rf_tune)
rf_tune$results %>% filter(mtry == as.numeric(rf_tune$bestTune))
defaultSummary(data.frame(obs = test_y,pred = predict(rf_tune, test_x)))
```

Optimum parametrelerle hesaplanmış RMSE, MAE ve R^2 değerleri Tablo 4.2'de gösterilmektedir. Burada tek değişkenle en iyi tahminlemeyi yapabileceğimizi yorumlayabiliriz. Şekil 4.13'te rüzgar hızına bağlı olarak gerçekleşen ve tahmin edilen değerlerin nasıl bir dağılıma sahip olduğu görülmektedir.

Tablo 4.2 Model ayarlama işleminden sonraki çıktılar

Mtry	**RMSE**	$\boldsymbol{R^2}$	**MAE**
1	0,05768	0,9573	0,02996
2	0,06135	0,9518	0,03148



Şekil 4.13 Rassal ormanlar regresyonu kullanılırken rüzgar hızı (m/s) ile türbinin güç üretimi (kW) arasındaki ilişkiyi gösteren dağılım grafiği

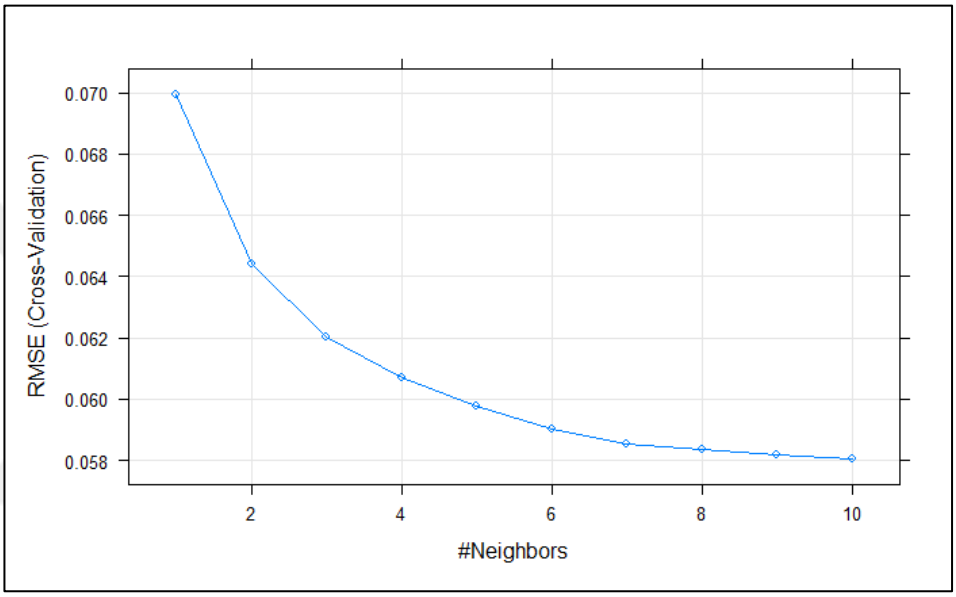
Singh vd. (2021) çalışmasında teorik güç, aktif güç, rüzgar hızı ve rüzgar yönü girdileri ile rassal ormanlar regresyonu kullanarak elde ettikleri performans ölçütleri Şekil 4.14'te gösterilmiştir. RF algoritmasının güç değerlerini yüksek performansla tahmin edildiği görülmüştür ancak performansı karar ağaçları (DT) regresyon modelinden daha iyi olmasına rağmen yüksek rüzgar hızı değerlerinde bu algoritmanın doğru tahminler üretemediği görülmüştür.

Regression Models	Performance Evaluation on Training Dataset					Performance Evaluation on Testing Dataset				
	MAE	MAPE	RMSE	MSE	R^2	MAE	MAPE	RMSE	MSE	R^2
Random Forest	0.0186	0.2966	0.0588	0.0040	0.9888	0.0277	0.3310	0.0672	0.0045	0.9651
K-NN	0.0278	0.2960	0.0580	0.0036	0.9742	0.0286	0.3248	0.0667	0.0044	0.9656
GBM	0.0260	0.0555	0.0228	0.0031	0.9897	0.0264	0.3012	0.0634	0.0040	0.9690
Decision Tree	0.0325	0.3213	0.0592	0.0055	0.9660	0.0336	0.3349	0.0884	0.0078	0.9497
Extra Tree	0.0274	0.2915	0.0522	0.0036	0.9782	0.0276	0.3243	0.0655	0.0041	0.9678

Şekil 4.14 Singh vd.(2021) rassal ormanlar regresyonu için çalışmasında elde ettiği bulgular

4.5 K-en Yakın Komşuluk (K-NN) Regresyonu için Bulgular

K-en yakın komşuluk regresyonu için öncelikle k parametresinin kaç olacağına dair denemeler yapılmış ve daha sonra modeli ayarlayarak en iyi k değerinin bulunması sağlanmıştır. Optimum k sayısı RMSE yani hata kareler oranına bağlı olarak değişen Şekil 4.15'ter yer alan grafikte açıkça görülmektedir. k=2 iken ve optimum seçilen k değerini gösteren değerler Tablo 4.3'teki gibidir.

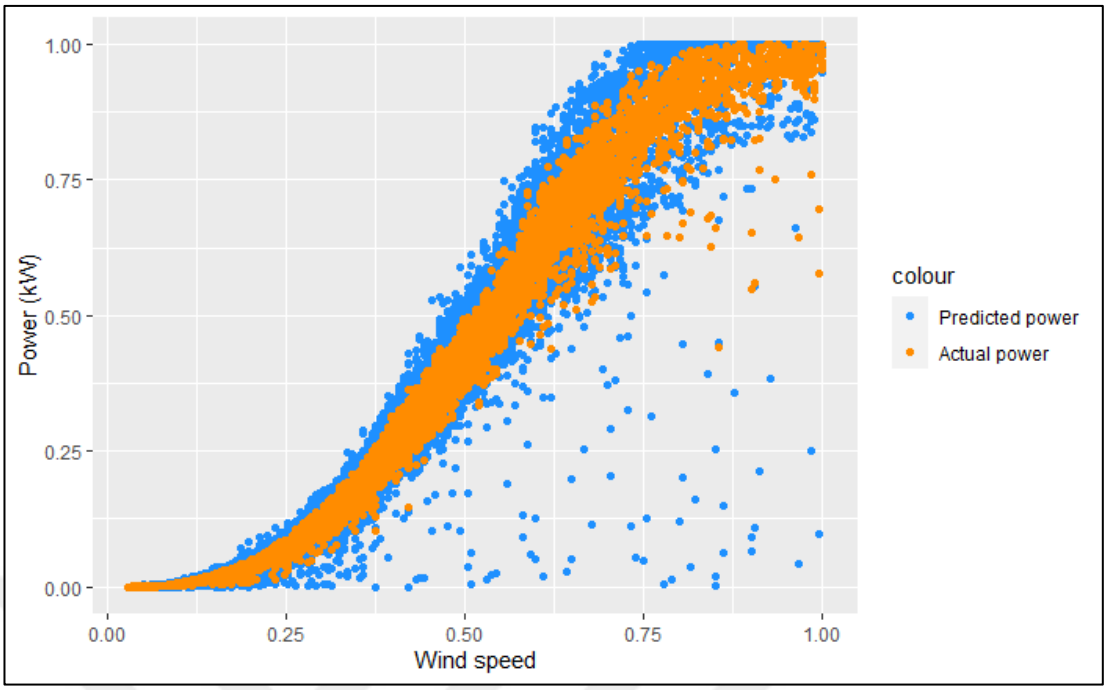


Şekil 4.15 Optimum k sayısının belirlenmesi

Tablo 4.3 k=2 iken ve model ayarlama işleminden sonraki çıktılar

k=2	**RMSE**	R^2	**MAE**
	0,06655	0,94280	0,03379
k=10	**RMSE**	R^2	**MAE**
	0,05637	0,95858	0,02890

Şekil 4.16'da nihai k-NN modeli kullanılarak rüzgar hızına bağlı olarak rüzgar gücünün tahmin ve gerçekleşen verilerinin arasındaki ilişkiyi gösteren dağılım grafiği bulunmaktadır.



Şekil 4.16 K-NN regresyonu kullanılırken rüzgar hızı (m/s) ile türbinin güç üretimi (kW) arasındaki ilişkiyi gösteren dağılım grafiği

Singh vd. (2021) çalışmasında teorik güç, aktif güç, rüzgar hızı ve rüzgar yönü girdileri ile k-en yakın komşuluk regresyonu kullanarak elde ettikleri performans ölçütleri Şekil 4.17'de gösterilmiştir. k-NN modelinde tahmin edilen güç değerlerinin çoğu gerçek ortalama güce veya aktif güce yakın çıkmıştır. Bundan dolayı, k-NN regresyon modelinin de tatmin edici bir performans gösterdiği görülmektedir.

Regression Models	Performance Evaluation on Training Dataset					Performance Evaluation on Testing Dataset				
	MAE	MAPE	RMSE	MSE	R^2	MAE	MAPE	RMSE	MSE	R^2
Random Forest	0.0186	0.2966	0.0588	0.0040	0.9888	0.0277	0.3310	0.0672	0.0045	0.9651
K-NN	0.0278	0.2960	0.0580	0.0036	0.9742	0.0286	0.3248	0.0667	0.0044	0.9656
GBM	0.0260	0.0555	0.0228	0.0031	0.9897	0.0264	0.3012	0.0634	0.0040	0.9690
Decision Tree	0.0325	0.3213	0.0592	0.0055	0.9660	0.0336	0.3349	0.0884	0.0078	0.9497
Extra Tree	0.0274	0.2915	0.0522	0.0036	0.9782	0.0276	0.3243	0.0655	0.0041	0.9678

Şekil 4.17 Singh vd.(2021) k-en yakın komşuluk regresyonu için çalışmasında elde ettiği bulgular

4.6 Destek Vektör Regresyonu (SVR) için Bulgular

Destek vektör regresyonunda optimize edilmesi gereken bazı parametreler bulunmaktadır. Bu parametreler cost, epsilon, kernel çekirdeği, gamma, vektör sayısı

olarak bilinmektedir. R programlama dilinde SVM elde edebilmek için kullanılan varsayılan argümanları aşağıda verilmiştir.

svm(x, y = NULL, scale = TRUE, type = NULL, kernel = “radial”, degree = 3, gamma = if (is.vector(x)) 1 else 1 / ncol(x), coef0 = 0, cost = 1, nu = 0.5, class.weights = NULL, cachesize = 40, tolerance = 0.001, epsilon = 0.1, shrinking = TRUE, cross = 0, probability = FALSE, fitted = TRUE, ..., subset, na.action = na.omit)

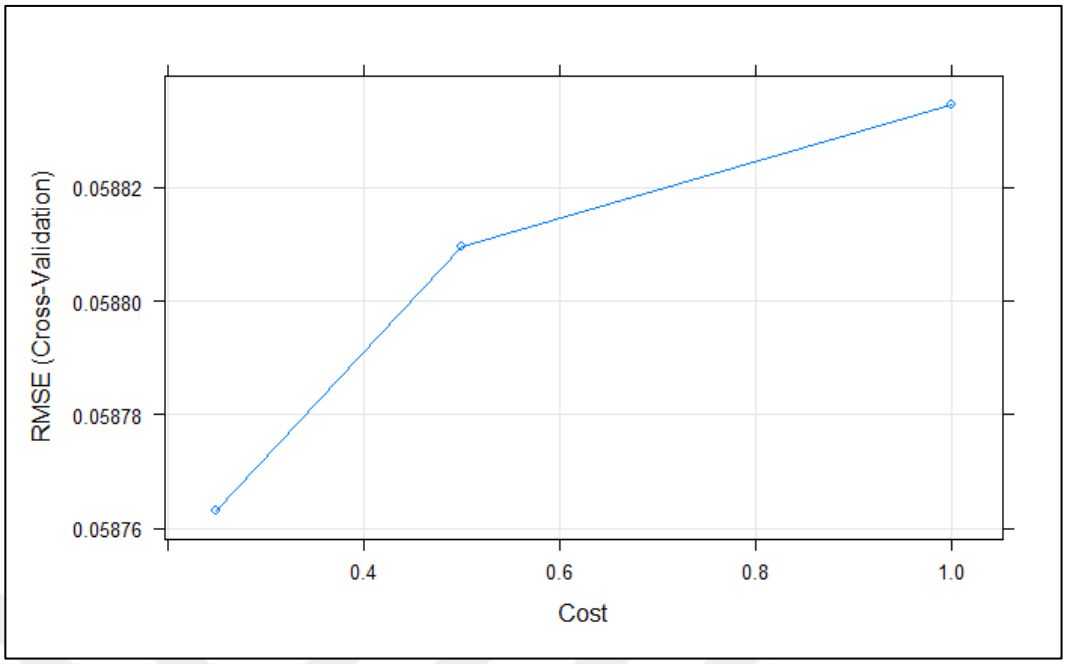
type : C-classification, nu-classification, one-classification (for novelty detection), eps-regression, nu-regression şeklinde tanımlanabilir. Bağımlı değişkenin (y) türüne göre ve kullanılacak amaca göre seçilir.

kernel: Eğitim ve tahminlemede kullanılır. Kullanılan dört farklı kernel fonksiyonu vardır.

gamma, **degree** ve **coef0** kernel belirlemede kullanılan parametrelerdir.

cost : Eğitim hatalarını ve marjları kontrol eden bir maliyet işlevine sahiptir. Marjı maksimize etmek için bir maliyet ortaya çıkmaktadır. Bu yüzden optimum çözümü belirlemek için kullanılır (default=1). Maliyet azaldıkça marj büyür ve sınıflandırma hataları ortaya çıkar. Bu çalışmada optimum maliyet Şekil 4.18’deki grafikte görülmektedir.

Modeli ayarlamadan önceki hata oranları ve açıklayıcılık oranları Tablo 4.4’te görülmektedir.

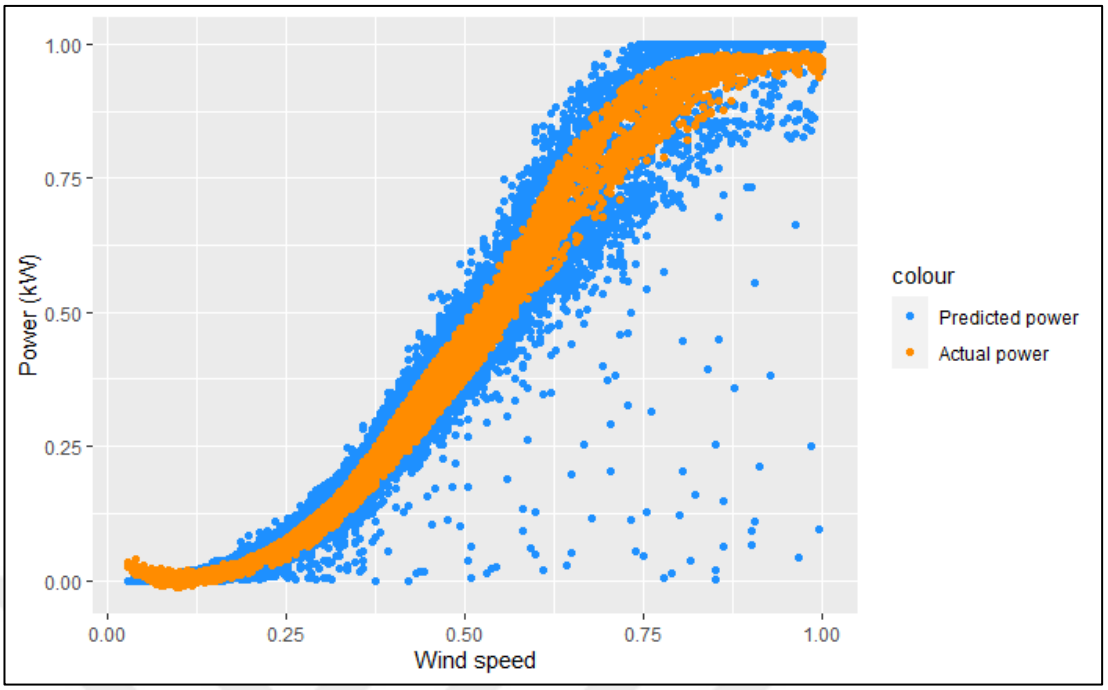


Şekil 4.18 Optimum cost (maliyetin) belirlenmesi

Tablo 4.4 Model svm ve model ayarlama işleminden sonraki çıktılar

Model svm (cost = 1 , epsilon = 0.1)	**RMSE**	R^2	**MAE**
	0,05804	0,95645	0,02993
Model tune svm (cost = 0.25, epsilon = 0.1)	**RMSE**	R^2	**MAE**
	0,05792	0,95675	0,02909

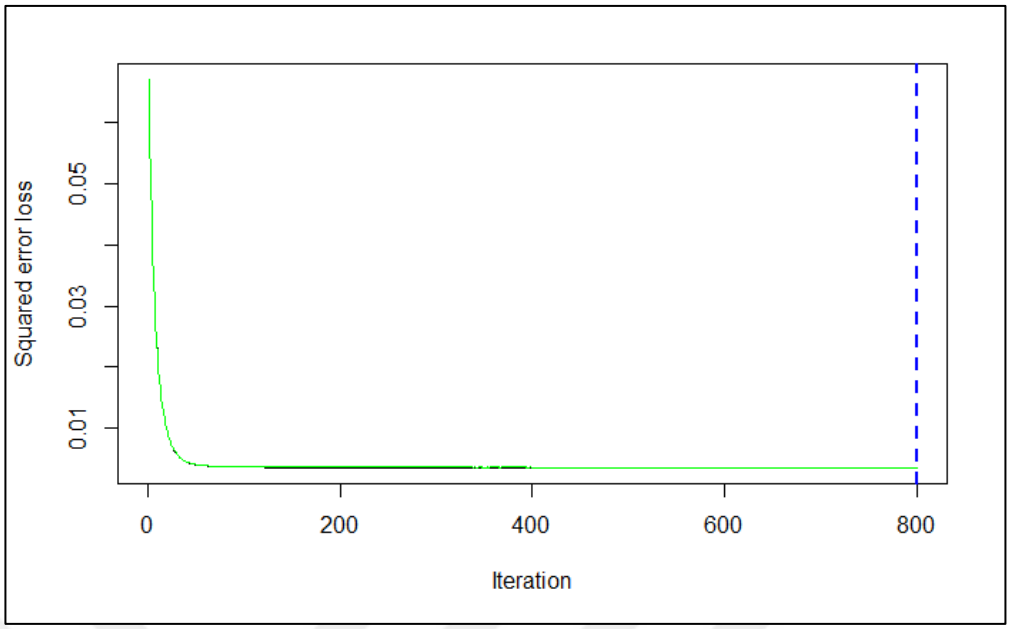
Şekil 4.19'da nihai SVR modeli kullanılarak rüzgar hızına bağlı olarak rüzgar gücünün tahmin ve gerçekleşen verilerinin arasındaki ilişkiyi gösteren dağılım grafiği bulunmaktadır.



Şekil 4.19 Destek vektör regresyonu kullanılırken rüzgar hızı (m/s) ile türbinin güç üretimi (kW) arasındaki ilişkiyi gösteren dağılım grafiği

4.7 Gradyan Arttırma Makinesi (GBM) için Bulgular

Gradyan arttırma makinelerinde kullanılan hiperparametreler arasında ağaç sayısı (ntrees), derinlik (interaction depth), öğrenme oranı (shrinkage) gibi bu parametrelerin model için optimize edilmeye ihtiyaçları vardır. Tablo 4.5'te varsayılan ve optimize edilmiş şekilde kıyaslanmış değerler bulunmaktadır. Şekil 4.20'de görüldüğü gibi, iterasyon sayısı yani ağaç sayısı arttıkça hata oranı giderek düşmektedir.

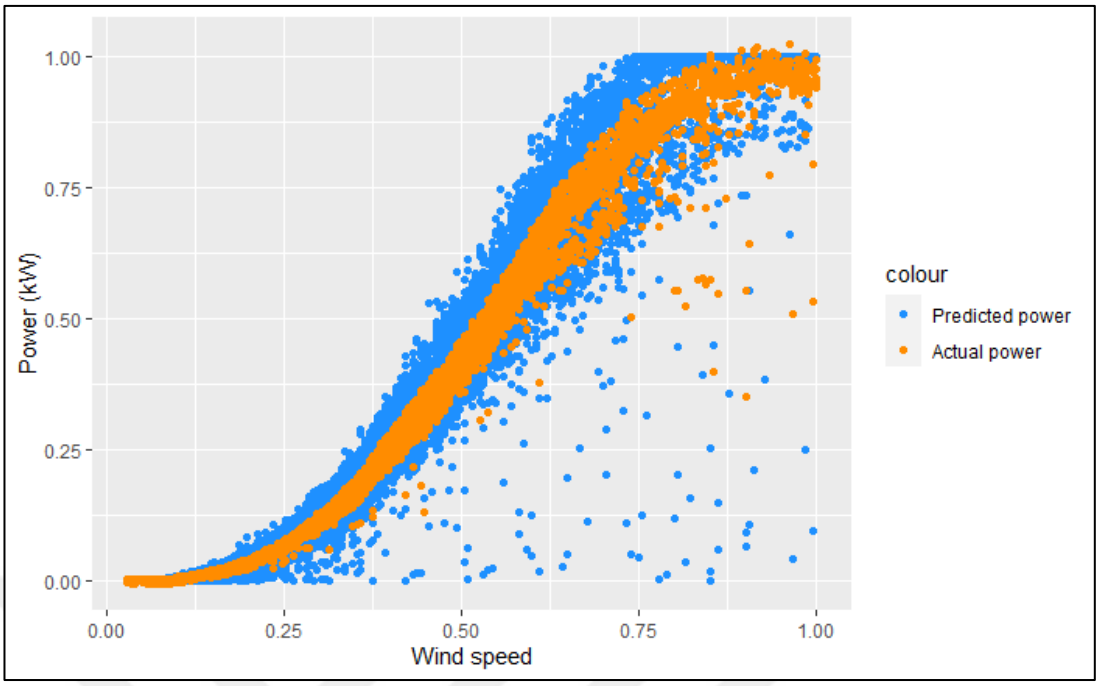


Şekil 4.20 Ağaç sayısının belirlenmesi

Tablo 4.5 Model GBM ve model ayarlama işleminden sonraki çıktılar

Model GBM (ntrees= 800, interaction depth =1, shrinkage= 0.1, cv.folds=10)	**RMSE**	R^2	**MAE**
	0,05874	0,9550	0,03263
Model GBM tune (ntrees= 150, interaction depth =9, shrinkage= 0.1, cv.folds=10)	**RMSE**	R^2	**MAE**
	0,05620	0,95957	0,02941

Şekil 4.21'de nihai GBM modeli kullanılarak rüzgar hızına bağlı olarak rüzgar gücünün tahmin ve gerçekleşen verilerinin arasındaki ilişkiyi gösteren dağılım grafiği bulunmaktadır.



Şekil 4.21 Gradyan arttırma makineleri kullanılırken rüzgar hızı (m/s) ile türbinin güç üretimi (kW) arasındaki ilişkiyi gösteren dağılım grafiği

Singh vd. (2021) çalışmasında teorik güç, aktif güç, rüzgar hızı ve rüzgar yönü girdileri ile GBM kullanarak elde ettikleri performans ölçütleri Şekil 4.22'de gösterilmiştir. Singh vd. (2021) çalışmasına göre, GBM karar ağaçları modelleri arasında tahmin performansı açısından daha iyi çıktılara sahiptir. Gradyan artırmada, bir makine öğrenimi modelinin performansını artırmak için bir grup zayıf öğrenici kullanılır. Zayıf öğrenenler genellikle karar ağaçlarıdır. Birleştirildiğinde, çıktıları daha iyi modellerle sonuçlanır.

Regression Models	Performance Evaluation on Training Dataset					Performance Evaluation on Testing Dataset				
	MAE	MAPE	RMSE	MSE	R^2	MAE	MAPE	RMSE	MSE	R^2
Random Forest	0.0186	0.2966	0.0588	0.0040	0.9888	0.0277	0.3310	0.0672	0.0045	0.9651
K-NN	0.0278	0.2960	0.0580	0.0036	0.9742	0.0286	0.3248	0.0667	0.0044	0.9656
GBM	0.0260	0.0555	0.0228	0.0031	0.9897	0.0264	0.3012	0.0634	0.0040	0.9690
Decision Tree	0.0325	0.3213	0.0592	0.0055	0.9660	0.0336	0.3349	0.0884	0.0078	0.9497
Extra Tree	0.0274	0.2915	0.0522	0.0036	0.9782	0.0276	0.3243	0.0655	0.0041	0.9678

Şekil 4.22 Singh vd.(2021) gradyan arttırma makinesi çalışmasında elde ettiği bulgular

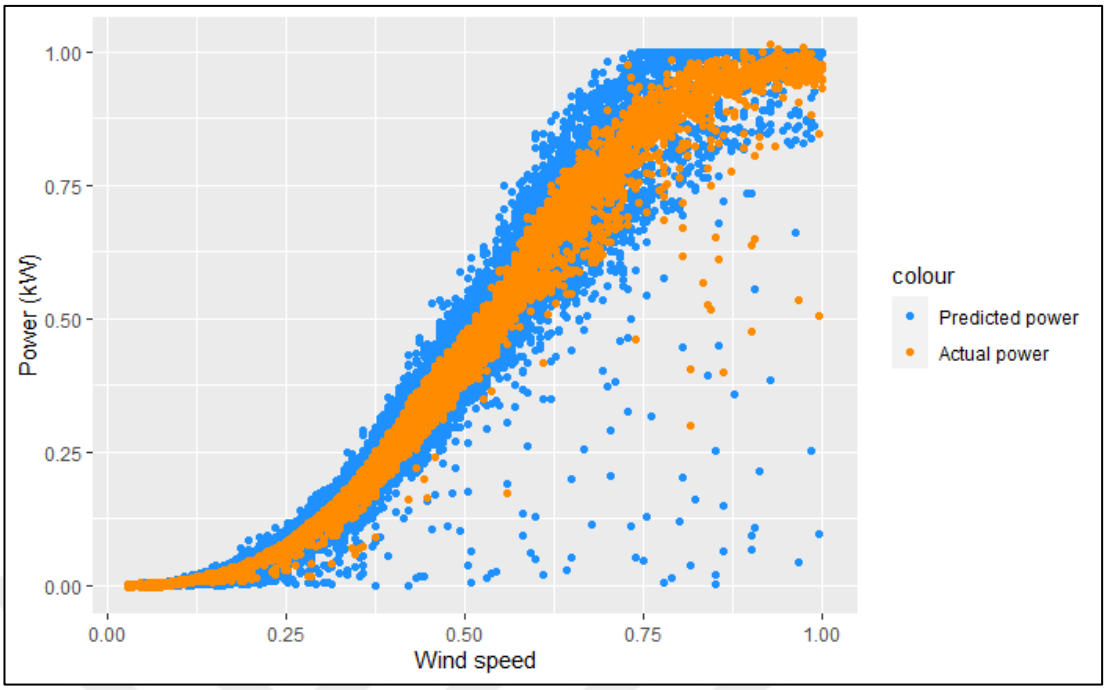
4.8 Aşırı Gradyan Arttırma (XGBoost) için Bulgular

XGBoost OpenMD çerçevesi sayesinde otomatik olarak paralel hesaplama yapmaktadır. Böylece klasik GBM'den 10 kat daha hızlı çalışır. Hız ve performans açısından GBM'den daha öndedir. Çok büyük veri setlerinde bile paralel mimarisi sayesinde hızdan faydalanmaktadır. Makine öğrenmesi dünyasında populer, aşırı öğrenmeye karşı duyarlı, sınıflandırma ve regresyon problemlerinde uygulanabilen bir algoritmadır. Örneklendirecek olursak, XGBoost seyreklik matrisinde sıfır gözlemlerini RAM'de tutmayarak hız imkanı sağlamaktadır. Girdi tipleri olarak; yoğunluk matrisi (R'ın yoğunluk matrisi), seyrek matrisi (R'ın seyreklik matrisi) ve veri sınıfı (xgb.DMatrix) olarak gözlenir. Seyreklik, regresyonda ya da sınıflandırma problemleri için seyrek girdileri kabul ederek buna göre optimize edilmiştir. Bazı parametrelerin optimizasyonuna ihtiyaç duyulabilir, bunlardan bazıları şunlardır; max.depth (ağaç derinliği), eta (öğrenme oranı), nthread (paralel olarak çalıştırılacak işlemci sayısı), nrounds (iterasyon sayısı ya da ağaç sayısı). Tablo 4.6'da bu parametrelerin optimize edilmeden önce ve sonraki değerleri gösterilmiştir.

Tablo 4.6 Model XGBoost ve model ayarlama işleminden sonraki çıktılar

Model XGBoost (max.depth = 4, eta=0.1, nthread=2, nrounds= 1000)	**RMSE**	R^2	**MAE**
	0,08127	0,91386	0,06029
Model XGBoost tune (alpha=1, eta=0, lambda=2, nrounds= 1000)	**RMSE**	R^2	**MAE**
	0,05518	0,96030	0,02840

Şekil 4.23'te nihai XGBoost modeli kullanılarak rüzgar hızına bağlı olarak rüzgar gücünün tahmin ve gerçekleşen verilerinin arasındaki ilişkiyi gösteren dağılım grafiği bulunmaktadır.



Şekil 4.23 XGBoost kullanılırken rüzgar hızı (m/s) ile türbinin güç üretimi (kW) arasındaki ilişkiyi gösteren dağılım grafiği

BÖLÜM 5
SONUÇ

Dünyada rüzgar ve güneş enerjisi gibi yenilenebilir enerji kaynakları giderek daha fazla kullandığından, bu tür enerji üretimini tahmin etmek gelecek için önemli hale gelmektedir. Rüzgar enerjisi bilindiği üzere alternatif enerji kaynağı olarak büyük ölçekte kullanılmaktadır. Rüzgar enerjisinin dalgalı doğası nedeniyle tahminleme süreçlerinde bazı zorluklarla karşılaşılmaktadır. Hata miktarının daha az olduğu doğru modeller elde etmek için farklı modeller üzerinde çalışmalar yapılmaktadır. Bu çalışmada rüzgar hızı ve rüzgar yönü girdilerine dayanarak rüzgar gücünü tahmin etmek için çeşitli makine öğrenmesi yöntemlerinin karşılaştırmalı analizi yapılmıştır. Bu karşılaştırmalı çalışmada özellikle k-NN, SVR, RF, GBM ve XGBoost modelleri ele alınmıştır. Bu amaca ulaşmak için Türkiye'nin İç Anadolu bölgesinde yer alan bir rüzgar santrali vaka çalışması olarak kullanılmıştır. ML modellerini eğitmek ve test etmek için Ocak 2019'dan Aralık 2020'ye kadar 10 dakikalık SCADA verileri kullanılmıştır. ML modellerinin tahmin performansını değerlendirmek için farklı hata fonksiyonları kullanılmıştır. Tablo 5.1'de görüldüğü üzere tahmin performansları karşılaştırılıp en iyi modelin XGBoost olduğu görülmüştür. Bu çalışmanın rüzgar enerjisi sektöründeki teknik uygulamalara katkıda bulunacağı öngörülmektedir. Bölgeye özgü rüzgar hızına bağlı olarak çıktılar değerlendirilmiştir.

Tablo 5.1 Modellerin tahmin performansının karşılaştırılması

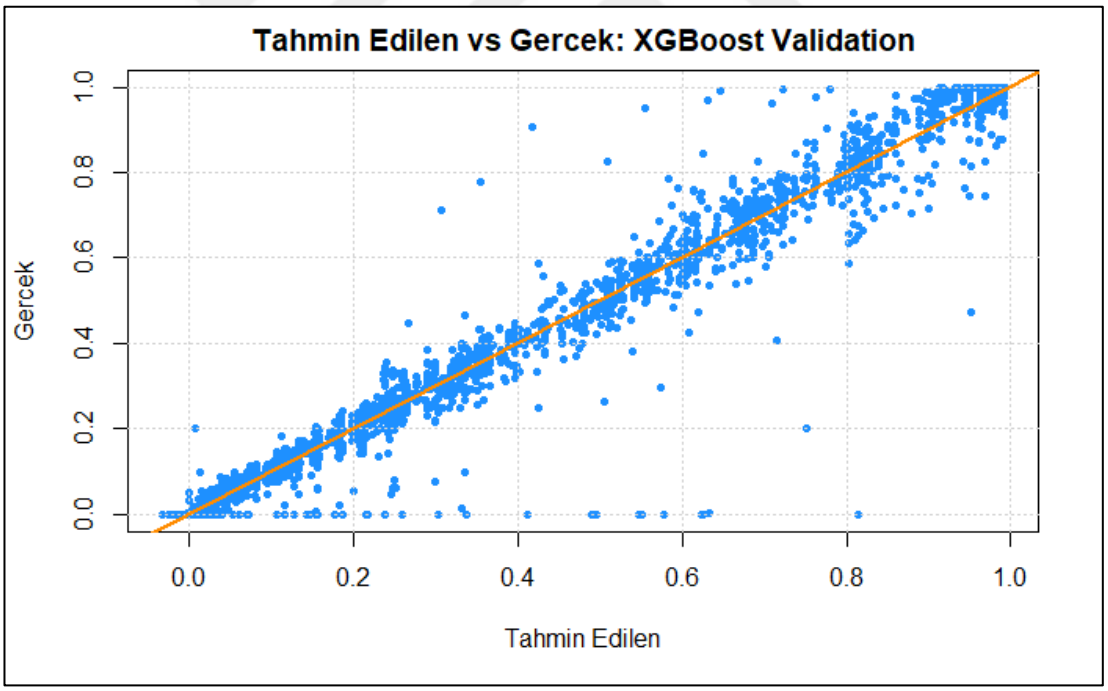
Modeller	**RMSE**	**MAE**	$\boldsymbol{R^2}$
RF	0,0556	0,0289	0,9596
k-NN	0,0571	0,0292	0,9574
SVR	0,0579	0,0291	0,9567
GBM	0,0555	0,0288	0,9597
XGBoost	0,0551	0,0284	0,9603

Çalışmanın sonucu olarak, en iyi modelin XGBoost olduğunu tespit ettikten sonra yapılan çalışmanın uygulanabilirliğini kontrol etmek için doğrulama yapılmıştır. Doğrulama testi için 2021 Ocak ayı 10 dakikalık verilerine çalışmada uyguladığımız yöntem olan CRISP-DM döngüsüne göre veri ön işleme, eğitim (%80) ve test (%20) verilerine ayırma işlemleri gerçekleştirilmiştir. Çalışmada kullanılan SCADA verileri

2019-2020 yıllarına ait 10 dakikalık kayıtlardır. Doğrulama testi için 2021 yılının Ocak ayı 10 dakikalık gerçekleşen değerlerle, kurduğumuz modeldeki tahmin değerlerini kıyaslayıp hata oranlarına bakılmıştır. Tablo 5.2'de görüldüğü gibi, kurulan modelde 2021 verileri aşırı öğrenme (overfit) olmadan tahmin edilmiştir ve model başarılı bir şekilde uygulanabilir olmuştur. Şekil 5.1'deki grafikte de görüldüğü gibi, doğrulama işlemi sonrası tahmin verilerinin 2021 verilerine karşı aşırı öğrenme olmadan gerçek verilere uyum sağlamıştır. Bu çalışmada bulunan öneriler gelecekteki çalışmalarda diğer makine öğrenmesi modellerinin geliştirilmesi, zamana bağlı tahminlerin geliştirilmesine öncülük edecektir.

Tablo 5.2 Doğrulama testi sonrası hata çıktıları

	RMSE	**MAE**	R^2
XGBoost	0,0551	0,0284	0,9603
Doğrulama	0,0579	0,0271	0,9701



Şekil 5.1 XGBoost doğrulama testi

KAYNAKLAR

Agarwal, D. (2021). *Introduction to SVM(Support Vector Machine) Along with Python Code*. Retrieved from Analytics Vidhya: https://www.analyticsvidhya.com/blog/2021/04/insight-into-svm-support-vector-machine-along-with-code/

Akpınar, H. (2014). *Data Veri Madenciliği Veri Analizi.* İstanbul: Papatya Yayıncılık Eğitim.

Alkesaiberi, A., Harrou, F., & Sun, Y. (2022). EfficientWind Power Prediction Using Machine Learning Methods: A Comparative Study. *Energies*, 2327.

Amit, Y., & Geman, D. (1997). Shape quantization and recognition with randomized trees. *Neural Computation*, 1545-1588.

Boser, B. E., Guyon, I. M., & Vapnik, V. N. (1992). A training algorithm for optimal margin classifiers. *In Proceedings of the fifth annual workshop on Computational learning theory (COLT '92)*, (pp. 144-152). New York. doi:https://doi.org/10.1145/130385.130401

Boulesteix, A.-L., Janitza, S., Kruppa, J., & König, I. R. (2012). Overview of random forest methodology and practical guidance with emphasis on computational biology and bioinformatics. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 493–507.

Breiman, L. (1996). Bagging Predictors. *Machine Learning*, 123–140.

Breiman, L. (2001). Random Forest. In L. Breiman, *Machine Learning* (pp. 5–32). Netherlands: Kluwer Academic.

Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (1999). *Classification and Regression Trees.*

Buturache, A.-N., & Stancu, S. (2021). Wind Energy Prediction Using Machine Learning. *Low Carbon Economy*(12), 1-21.

Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. *In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '16)* (pp. 785-794). New York: Association for Computing Machinery.

Cover, T. (1968). Rates of convergence for nearest neighbor procedures. *Proceedings of the Hawaii International Conference on Systems Sciences*, (pp. 413-415).

Cover, T., & Hart, P. E. (1967). Nearest neighbor pattern classification. *IEEE Transactions on Information Theory*, 21-27. doi:10.1109/TIT.1967.1053964

Cutler, D. R., Jr., T. C., Beard, K. H., Cutler, A., Hess, K. T., Gibson, J., & Lawler, J. J. (2007). Random forests for classification in ecology. *Ecology*, 2783–2792. doi:10.1890/07-0539.1

Delgado, I., & Fahim, M. (2021). Wind Turbine Data Analysis and LSTM-Based Prediction in SCADA System. *Energies*(14), 125. doi:10.3390/en14010125

Dhiman, H. S., & Deb, D. (2016). A Review of Wİnd Speed & Wind Power Forecasting Techniques. *In Proceedings of the 2016 International Conference on Probabilistic Methods Applied to Power Systems (PMAPS)*, (pp. 1-7). Beijing, China.

Dietterich, T. (1999). An experimental comparison of three methods for constructing ensembles of decision trees:Bagging, boosting and randomization. *Machine Learning*, 1-22.

Ekelik, H., & Altaş, D. (2019). Dijital Reklam Verilerinden Yararlanarak Potansiyel Konut Alıcılarının Rastgele Orman Yöntemiyle Sınıflandırılması. *İktisat Araştırmaları Dergisi*, 28-45.

ENERCON. (2021, Şubat). *ENERCON product overview*. Retrieved from https://www.enercon.de/fileadmin/Redakteur/Medien-Portal/broschueren/pdf/en/ENERCON_Produkt_en_06_2015.pdf

Eyecioğlu, Ö., Hangun, B., Kayışlı, K., & Yeşilbudak, M. (2019). Performance Comparison of Different Machine Learning Algorithms on the Prediction of Wind Turbine Power Generation. *8th International Conference on Renewable Energy Research and Applications*, (pp. 3-6). Brasov.

Fawagreh, K., Gaber, M. M., & Elyan, E. (2014). Random forests: from early developments to recent advancements. *Systems Science & Control Engineering: An Open Access Journal*, 602-609. doi:10.1080/21642583.2014.956265

Flaxman, A. D., Vahdatpour, A., Green, S., James, S. L., & Murray, C. J. (2011). Random forests for verbal autopsy analysis: Multisite validation study using clinical diagnostic gold standards. *Population Health Metrics*, 1-11.

Freund, Y., & Schapire, R. E. (1999). Adaptive Game Playing Using Multiplicative Weights. *Games and Economic Behavior*, 79-103. doi:https://doi.org/10.1006/game.1999.0738

Fugon, L., Juban, J., & Kariniotakis, G. (2008). Data mining for wind power forecasting. *European Wind Energy Conference.* Belgium: European Wind Energy Conference & Exhibition EWEC 2008, Brussels : Belgium (2008).

Gao, D., Zhang, Y.-X., & Zhao, Y.-H. (2009). Random forest algorithm for classification of multiwavelength data. *Research in Astronomy and Astrophysics*, 14-39. doi:10.1088/1674-4527/9/2/011

Greenwell, B. M., & Boehmke, B. (2019). *Hands-On Machine Learning with R.* Chapman and Hall/CRC.

GWEC. (2023, Ocak). *Global Wind Report 2022*. Retrieved from https://gwec.net/wp-content/uploads/2022/03/GWEC-GLOBAL-WIND-REPORT-2022.pdf

Hastie, T., Tibshirani, R., & Friedman, J. (2001). *The Elements of Statistical Learning Data Mining, Inference and Prediction.* Springer.

Ho, T. K. (1998). The random subspace method for constructing decision forests. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 832-844.

Hu, W. (2009). Identifying predictive markers of chemosensitivity of breast cancer with random forests. *J. Biomedical Science and Engineering*, 59-64.

Huang, X., & Jiang, A. (2022). Wind Power Generation Forecast Based on Multi-Step Informer Network. *Energies*. doi:10.3390/en15186642

IBM Cloud Education. (2020, Ağustos). *What is Supervised Learning?* Retrieved from IBM: https://www.ibm.com/cloud/learn/supervised-learning#:~:text=Supervised%20learning%2C%20also%20known%20as,data%20or%20predict%20outcomes%20accurately

IBM Cloud Education. (2020, Eylül). *What is Unsupervised Learning?* Retrieved from IBM: https://www.ibm.com/cloud/learn/unsupervised-learning

İlkiliç, C. (2012). Wind energy and assessment of wind energy potential in Turkey. *Renewable and Sustainable Energy Reviews*(16), 1165-1173.

Işık, E., & Ünal, O. (2020). *Rüzgar Türbini Kanat Tasarımı ve Analizi.* Retrieved from Research Gate :

https://www.researchgate.net/publication/339956087_Ruzgar_Turbini_Kanat_Tasarimi_ve_Analizi

Karabağ, N., Kayıkcı, C. B., & Öngen, A. (2021). %100 Yenilenebilir Enerjiye Geçiş Yolunda Dünya ve Türkiye. *Avrupa Bilim ve Teknoloji Dergisi*, 230-240.

Kıvrak, M. (2021). *Sınıflandırma Problemlerinde Topluluk Öğrenme Yöntemlerinin İncelenmesi ve Küçük Hücreli Dışı Akciğer Kanseri Verileri Üzerine Bir Uygulama.* Malatya: İnönü Üniversitesi.

Klassen, M., Cummings, M., & Saldaña-González, G. (2008). Investigation of random forest performance with cancer microarray data. *Proceedings of the ISCA 23rd International Conference on Computers and Their Applications*, (pp. 64-69). Mexico.

Kuhn, M., & Johnson, K. (2013). *Applied Predictive Modeling.* New York: Springer.

Kusiak, A., Zheng, H., & Song, Z. (2009). Short-Term Prediction of Wind Farm Power: A Data Mining Approach. *IEEE Transactions on Energy Conversion*, 125-136.

Larose, D. T. (2005). What Is Data Mining? In D. T. Larose, *Discovering knowledge in data - An Introduction to Data Mining.* A JOHN WILEY & SONS, INC.

Löw, F., Schorcht, G., Michel, U., & Conrad, C. (2012). Per-field crop classification in irrigated agricultural regions in Middle Asia using random forest and support vector machine ensemble. *In SPIE remote sensing.* Edinburg, United Kingdom.

Morde, V. (2019). *XGBoost Algorithm: Long May She Reign!* Retrieved from Towards Data Science: https://towardsdatascience.com/https-medium-com-vishalmorde-xgboost-algorithm-long-she-may-rein-edd9f99be63d

Özbay, Ö. (2015). Veri Madenciliği Kavramları ve Eğitimde Veri Madenciliği Uygulamaları. *Uluslararsı Eğitim Bilimleri Dergisi*, 262-272.

Özen, M. S. (2021). *Rüzgar Enerjisi Sektörü ve İzmir Denizüstü Rüzgar Enerjisi Yol Haritası.* İzmir: İzmir Kalkınma Ajansı. Retrieved Şubat 2021, from https://izka.org.tr/wp-content/uploads/2021/08/ruzgar_enerjisi_sekto%CC%88ru%CC%88-izmir_denizustu_ruzgar_enerjisi_yol-haritasi-2.pdf

Patle, A., & Chouhan, D. S. (2013). SVM kernel functions for classification. *2013 International Conference on Advances in Technology and Engineering (ICATE)* (pp. 1-9). Mumbai, India: IEEE.

Rodriguez-Galiano, V. F., Sanchez-Castillo, M., DAsh, J., Atkinson, P. M., & Ojeda-Zujar, J. (2016). Modelling interannual variation in the spring and autumn land surface phenology of the European forest. *Biogeosciences*, 3305-3317. doi:10.5194/bg-13-3305-2016

Shabbir, N., Kütt, L., Jawad, M., Husev, O., Rehman, A. U., Gardezi, A., . . . Choi, J.-G. (2021). Short-Term Wind Energy Forecasting Using Deep Learning-Based Predictive Analytics. *Computers, Materials & Continua*, 1017–1033.

Singh , U., Rizwan, M., Alaraj, M., & Alsaidan, I. (2021). A Machine Learning-Based Gradient Boosting Regression Approach for Wind Power Production Forecasting: A Step towards Smart Grid Environments. *Energies*, *14*(16), 5196.

Singh, H. (2022). *Understanding Gradient Boosting Machines*. Retrieved from Published in: https://towardsdatascience.com/understanding-gradient-boosting-machines-9be756fe76ab

Singh, U., & Rizwan, M. (2022). SCADA system dataset exploration and machine learning based forecast for wind turbines. *Elsevier BV*, 100640.

Son, N., Yang, S., & Na, J. (2019). Hybrid Forecasting Model for Short-Term Wind Power Prediction Using Modified Long Short-Term Memory. *Energies*(12), 3901. doi:10.3390/en12203901

Sun, J., Du, W., & Shi, N. (2018). A Survey of kNN Algorithm. *Information Engineering and Applied Computing*.

Taşcı, M. (2020). Uçuş Verilerinin Kaba Küme Yöntemi Kullanılarak İndirgenmesi. *EJONS International Journal on Mathematic, Engineering and Natural Sciences* , 625-637.

Taunk, K., De, S., Verma, S., & Swetapadma, A. (2019). A Brief Review of Nearest Neighbor Algorithm for Learning and Classification. *2019 International Conference on Intelligent Computing and Control Systems (ICCS).* doi:10.1109/ICCS45141.2019.9065747

TEİAŞ. (2022). *Gerçek Zamanlı Üretim - Gerçekleşen Üretim*. Retrieved Temmuz 2022, from EPİAŞ Şeffaflık Paltformu:

https://seffaflik.epias.com.tr/transparency/uretim/gerceklesen-uretim/gercek-zamanli-uretim.xhtml

Teixeira-Pinto, A. (2022). *Machine Learning for Biostatistics*. Retrieved from K-nearest Neighbours Regression: https://bookdown.org/tpinto_home/Regression-and-Classification/k-nearest-neighbours-regression.html

TÜREB. (2023, Ocak). *Rüzgar Enerjisi Santralleri Raporu.* Retrieved from TÜREB: Türkiye Rüzgar Enerjisi Birliği: https://app.powerbi.com/view?r=eyJrIjoiNmFmYWY0MTYtNjUyNS00NzQ1LWIwMTMtOTI5ZTNkM2FlYWIxIiwidCI6ImU5YzY0NjU4LWFkMWQtNDUwOS1hODk0LTE2NWZhYjU2NjEyMyIsImMiOjl9

Vontela, S. (2022, Mayıs). *Building KNN Regression Algorithm from Scratch*. Retrieved from medium.com: https://medium.com/where-quant-meets-data-science/building-k-nearest-neighbour-algorithm-from-scratch-bd0c5df13192

Yürek, Ö. E., Birant, D., & Yürek, İ. (2021). Wind Power Generation Prediction Using Machine Learning Algorithms. *Dokuz Eylül University Faculty of Engineering Journal of Science and Engineering, 23*(67), 107-119.

Zaklouta, F., Stanciulescu, B., & Hamdoun, O. (2011). Traffic sign classification using kd trees and random forests. *In The 2011 international joint conference on neural networks (IJCNN)*, (pp. 2151–2155). San Jose.