MAKİNE ÖĞRENMESİNDE DEĞİŞKEN SEÇİM YÖNTEMLERİNİN KARŞILAŞTIRILMASI: EV ENERJİSİ TÜKETİM TAHMİNİ

COMPARISON OF VARIABLE SELECTION IN MACHINE LEARNING METHODS: HOUSEHOLD ENERGY CONSUMPTION ESTIMATION

NURİ BERK URAL

PROF. DR. MERAL ÇETİN

Tez Danışmanı

Hacettepe Üniversitesi

Lisansüstü Eğitim-Öğretim ve Sınav Yönetmeliğinin

İstatistik Anabilim Dalı için Öngördüğü

YÜKSEK LİSANS TEZİ olarak hazırlanmıştır.

2024

ÖZET

MAKİNE ÖĞRENMESİNDE DEĞİŞKEN SEÇİM YÖNTEMLERİNİN KARŞILAŞTIRILMASI: EV ENERJİSİ TÜKETİM TAHMİNİ

Nuri Berk URAL

Yüksek Lisans, İstatistik Bölümü

Tez Danışmanı: Prof. Dr. Meral ÇETİN

Haziran 2024, 80 Sayfa

Günümüz dijital çağında, her geçen gün artan dijital aktiviteler ve teknolojik gelişmeler sayesinde üretilen veri miktarı hızla büyümekte ve bu durum, "büyük veri" olarak adlandırılan yeni bir çalışma alanının doğuşuna zemin hazırlamaktadır. Büyük veri kavramı, sadece hacmiyle değil, çeşitliliği ve üretim hızıyla da geleneksel veri işleme tekniklerinin ötesine geçmektedir. Geleneksel istatistiksel yöntemler, bu verinin karmaşıklığı ve büyüklüğü karşısında yetersiz kalmaktadır. Bu nedenle, veri bilimi disiplini içerisinde bu devasa veri akışını etkili bir şekilde kontrol edebilmek, analiz edebilmek ve değerli bilgilere dönüştürebilmek için yeni ve daha gelişmiş yöntemlerin ve teknolojilerin geliştirilmesi kaçınılmaz hale gelmiştir.

Bu yeni yöntemler, makine öğrenmesi ve yapay zekâ gibi alanlarda da önemli ilerlemelere yol açarak veriden anlam çıkarma süreçlerini daha etkin ve verimli hale getirmiştir. Bu durum, veri bilimi alanının sadece akademik bir merak konusu olmaktan çıkıp iş dünyası, sağlık, finans ve birçok diğer sektörde stratejik karar alma süreçlerinde kritik bir role sahip olmasına neden

olmuştur. Bu gelişmelerle birlikte model oluşturma süreci de çok daha karmaşık hale gelmiştir. Bu noktada, modelin doğru tahmin performansını arttırmak ve anlamlı sonuçlar elde etmek için değişken seçiminin ne kadar kritik olduğu ortaya çıkmaktadır. Yanlış değişken seçimi, modelin tahmin performansını olumsuz yönde etkileyebilir ve yanıltıcı sonuçların ortaya çıkmasına zemin hazırlayabilir.

Değişken seçimi, büyük veri kümelerinden elde edilen anlamlı ve doğru sonuçlar için kritik bir adımdır. Yanlış veya önemsiz değişkenlerin seçimi, modelin genel tahmin kabiliyetini ciddi şekilde bozabilir, yanıltıcı sonuçlara yol açabilir ve yanlış kararların alınmasına sebep olabilir. Bu nedenle, veri bilimi pratiklerinde doğru değişkenleri belirleyebilmek için gelişmiş seçim teknikleri ve algoritmalarının kullanımı hayati öneme sahiptir. Bu teknikler, modelin karmaşıklığını yönetmeye, aşırı uyuma (overfitting) karşı korumaya ve en önemlisi tahmin performansını iyileştirmeye yardımcı olur. Özellikle Makine Öğrenmesi (ML) ve Yapay Zekâ (AI) modellerinde doğru değişken seçimi, modelin gerçek dünya verileri üzerindeki genelleme kapasitesini artırarak daha güvenilir ve doğruluk oranı yüksek sonuçlar üretmesine olanak tanıyabilir.

Bu tez çalışması, enerji tüketimi tahmininde değişken seçim yöntemlerinin tahmin performansındaki rolünü incelemektedir. Bu kapsamda, çeşitli ML algoritmaları kullanılarak değişken seçim yöntemlerinin etkinliği ve bu yöntemlerle oluşturulan modellerin performansları karşılaştırılmıştır. Çalışmada kullanılan veri kümesi, ev aletlerinin enerji tüketimini tahmin etmek amacıyla oluşturulmuş bir veri kümesidir. Bu veri kümesi, bir evdeki çeşitli odalarda ve dış cephede yerleştirilen sensörlerle 4 buçuk ay boyunca her 10 dakikada bir alınan sıcaklık ve nem ölçümlerini içermektedir. Toplamda 19735 gözlem ve 28 değişkenden oluşmaktadır. Kayıp veya eksik gözlem bulunmamaktadır.

Çalışmanın temel amacı, değişken seçim yöntemlerinin ML algoritmalarının tahmin performansına olan etkilerini detaylı bir şekilde değerlendirmektir. Bu kapsamda, Korelasyon Tabanlı Seçim (CFS), Varyans Tabanlı Seçim, İleriye Doğru Seçim, Geriye Doğru Eleyerek Seçim, Adımsal Seçim, Genetik Algoritmalar Tabanlı Seçim, Lasso Regresyon Tabanlı Seçim, Ridge Regresyon Tabanlı Seçim ve Robust (Sağlam) Değişken Seçim Yöntemi kullanılmıştır.

Her bir değişken seçim yöntemi sonrası seçilen değişkenlerle Doğrusal Regresyon, Karar Ağaçları, Rastgele Ormanlar, Destek Vektör Makineleri, Temel Bileşenler Analizi ve Yapay Sinir Ağları algoritmaları kullanılarak modeller oluşturulmuş ve bu modellerin performansları Ortalama Mutlak Hata (MAE), Hata Kareler Ortalaması (MSE) ve Açıklanma Oranı (R^2) ölçütleri kullanılarak değerlendirilmiştir.

Çalışmanın sonuçları, farklı değişken seçim yöntemleri ve ML algoritmalarının enerji tüketimi tahmin performansı üzerindeki etkilerini karşılaştırmalı olarak sunmakta ve bu alanda yapılan diğer çalışmalarla paralellikler kurarak literatüre katkı sağlamaktadır. Özellikle, hangi değişken seçim yönteminin ve ML algoritmasının enerji tüketimi tahmini için en uygun olduğu konusunda önemli bulgular elde edilmiştir. Bu bulgular, veri bilimcileri ve araştırmacılar için veri kümelerine uygun yöntem ve algoritma seçiminde rehberlik edecek niteliktedir. Çalışma, veri bilimi alanında bilgi birikiminin artırılmasına, araştırma ve uygulama metodolojilerinin geliştirilmesine ve bu dinamik disiplinin ilerlemesine katkıda bulunmayı amaçlamaktadır.

Anahtar kelimeler: Değişken seçim yöntemleri, Makine öğrenmesi algoritmaları, Enerji tüketimi tahmini

ABSTRACT

COMPARISON OF VARIABLE SELECTION IN MACHINE LEARNING METHODS: HOUSEHOLD ENERGY CONSUMPTION ESTIMATION

Nuri Berk URAL

Master of Science, Department of Statistics

Supervisor: Prof. Dr. Meral ÇETİN

June 2024, 80 Pages

In today's digital age, the amount of data generated is rapidly increasing due to ever-growing digital activities and technological advancements, paving the way for a new field of study known as "big data." The concept of big data goes beyond traditional data processing techniques not only due to its volume but also because of its variety and velocity. Traditional statistical methods fall short in the face of the complexity and scale of this data. Therefore, within the discipline of data science, it has become inevitable to develop new and more advanced methods and technologies to effectively control, analyze, and transform this massive data flow into valuable insights.

These new methods have also led to significant advancements in fields such as Machine Learning and Artificial Intelligence, making data interpretation processes more efficient and effective. This evolution has positioned data science not merely as an academic curiosity but as a critical component in strategic decision-making processes in business, healthcare, finance, and many other sectors. Along with these developments, the model-building process has also become much more complex. At this point, the importance of variable selection to enhance model prediction performance and achieve meaningful results becomes evident. Incorrect

variable selection can negatively impact the model's prediction performance and lead to misleading results.

Variable selection is a critical step in obtaining meaningful and accurate results from large datasets. The selection of incorrect or irrelevant variables can severely degrade the overall predictive ability of the model, lead to misleading outcomes, and result in wrong decisions. Therefore, the use of advanced selection techniques and algorithms to identify the correct variables is of vital importance in data science practices. These techniques help manage model complexity, protect against overfitting, and most importantly, improve prediction performance. In particular, accurate variable selection in Machine Learning (ML) and Artificial Intelligence (AI) models can enhance the model's generalization capacity on real-world data, leading to more reliable and high-accuracy results.

This study examines the role of variable selection methods in prediction performance for energy consumption forecasting. Within this scope, the effectiveness of variable selection methods and the performance of models created using these methods are compared using various Machine Learning algorithms. The dataset used in the study is designed to predict the energy consumption of household appliances. This dataset includes temperature and humidity measurements taken every 10 minutes for 4.5 months by sensors placed in various rooms and outside the house. It consists of a total of 19,735 observations and 28 variables, with no missing or incomplete observations.

The primary objective of the study is to evaluate the detailed effects of variable selection methods on the prediction performance of machine learning algorithms. Within this scope, methods such as Correlation-Based Feature Selection (CFS), Variance-Based Selection, Forward Selection, Backward Elimination, Stepwise Selection, Genetic Algorithms-Based Selection, Lasso Regression-Based Selection, Ridge Regression-Based Selection and Robust Feature Selection Method were used. After each variable selection method, models were created using Linear Regression, Decision Trees, Random Forests, Support Vector Machines, Principal Component Analysis, and Neural Networks algorithms, and the performance of these models was evaluated using Mean Absolute Error (MAE), Mean Squared Error (MSE), and R^2 metrics.

The results of the study present a comparative analysis of the impact of different variable selection methods and machine learning algorithms on the performance of energy consumption prediction, contributing to the literature by drawing parallels with other studies in this field. Significant findings were obtained regarding which variable selection method and machine learning algorithm are most suitable for energy consumption forecasting. These findings provide guidance for data scientists and researchers in selecting appropriate methods and algorithms for their datasets. The study aims to contribute to the advancement of knowledge, research, and application methodologies in the dynamic discipline of data science.

Keywords: Variable selection methods, Machine learning algorithms, Energy consumption forecasting

TEŞEKKÜR

Tez çalışmam süresince bilgisi ve tecrübesiyle bana rehberlik eden, sabır ve anlayışını benden esirgemeyen danışman hocam Sayın Prof. Dr. Meral ÇETİN'e,

Değerli katkı ve önerileriyle tezime yön veren hocalarım Sayın Prof. Dr. Meral EBEGİL ve Dr. Öğr. Üyesi Onur TOKA'ya teşekkürlerimi sunarım.

Beni her daim koşulsuz destekleyen aileme; çalışırken sabote etme amacı olmadan klavyemin üzerinde uyuyan kedim Calypso ve beni her gördüğünde kucağıma atlayıp sevgisini hep hissettiren köpeğim 13'e sonsuz müteşekkirim.

Nuri Berk URAL

Haziran 2024, Adana

İÇİNDEKİLER

ÖZET i
ABSTRACT iv
TEŞEKKÜR vii
ÇİZELGELER LİSTESİ xi
ŞEKİLLER DİZİNİ xii
KISALTMALAR xiii
1. GİRİŞ 1
2. GENEL BİLGİLER 5
2.1. Makine Öğrenmesi 5
2.2. Makine Öğrenmesi Süreci 9
2.2.1. Problem Tanımı ve Hedef Belirleme 9
2.2.2. Veri Toplama 9
2.2.3. Veri Ön İşleme 10
2.2.4. Veri Görselleştirme 10
2.2.5. Değişken Seçimi 10
2.2.6. Model Değerlendirme 10
2.2.7. Model Seçimi ve Eğitimi 11
2.3. Makine Öğrenmesi Algoritmaları 11
2.3.1. Gözetimli Öğrenme Algoritmaları 11
2.3.1.1. Doğrusal Regresyon 12
2.3.1.2. Lojistik Regresyon 12
2.3.1.3. Karar Ağaçları 13
2.3.1.4. Rastgele Ormanlar 13

2.3.1.5. Destek Vektör Makineleri 14
2.3.1.6. Destek Vektör Regresyonu 15
2.3.1.7. Yapay Sinir Ağları 16
2.3.2. Gözetimsiz Öğrenme Algoritmaları 17
2.3.2.1. K-Ortalama Kümeleme 17
2.3.2.2. Hiyerarşik Kümeleme 17
2.3.2.3. Temel Bileşenler Analizi 19
2.3.2.4. Temel Bileşenler Regresyonu 20
2.3.2. Yarı-Gözetimli Öğrenme Algoritmaları 21
2.3.2.1. Kendini Eğitme 21
2.3.2.2. Eş-Eğitim 21
2.3.3. Pekiştirmeli Öğrenme Algoritmaları 21
2.4. Değişken Seçimi 22
2.5. Değişken Seçim Yöntemleri 23
2.5.1. Filtreleme Yöntemleri 23
2.5.1.1 Korelasyon Analizi Tabanlı Seçim 23
2.5.1.2. Ki-Kare Testi Tabanlı Seçim 24
2.5.1.3. Varyans Tabanlı Seçim 24
2.5.2. Sarmalama Yöntemleri 24
2.5.2.1. İleriye Doğru Seçim 25
2.5.2.2. Geriye Doğru Eleyerek Seçim 25
2.5.2.3. Adımsal Seçim 25
2.5.2.4. Genetik Algoritmalar Tabanlı Seçim 25
2.5.3. Gömülü Yöntemler 26
2.5.3.1. Lasso Regresyon Tabanlı Seçim 26
2.5.3.2. Ridge Regresyon Tabanlı Seçim 26
2.5.3.3. Karar Ağacı Tabanlı Seçim 27

2.5.4. Sağlam (Robust) Değişken Seçim Yöntemi 28
3. UYGULAMA 30
3.1. Veri Kümesi 30
3.2. Yöntem 39
3.3. Makine Öğrenmesi Algoritmaları Parametreleri 41
3.3.1. Doğrusal Regresyon Algoritması 42
3.3.2. Karar Ağaçları Algoritması 42
3.3.3. Rastgele Ormanlar Algoritması 43
3.3.4. Destek Vektör Makineleri Algoritması 44
3.3.5. Temel Bileşenler Analizi Algoritması 44
3.3.6. Yapay Sinir Ağları Algoritması 45
3.4. Model Kurulum ve Performansları 47
3.4.1. Korelasyon Tabanlı Seçim 47
3.4.2. Varyans Tabanlı Seçim 49
3.4.3. İleriye Doğru Seçim 52
3.4.4. Geriye Doğru Eleyerek Seçim 55
3.4.5. Adımsal Seçim 58
3.4.6. Genetik Algoritmalar Tabanlı Seçim 60
3.4.7. Lasso Regresyon Tabanlı Seçim 63
3.4.8. Ridge Regresyon Tabanlı Seçim 66
3.4.9. Sağlam Değişken Seçim Yöntemi 70
4. SONUÇ 73
5. KAYNAKLAR 76

ÇİZELGELER

Çizelge 3.1. Değişkenlerin açıklamaları ve birimleri

Çizelge 3.2. Kullanılan Yöntem, Algoritma ve Ölçütler

Çizelge 3.3. Korelasyon Tabanlı Seçim'e ilişkin sonuçlar

Çizelge 3.4. Varyans Tabanlı Seçim'e ilişkin sonuçlar

Çizelge 3.5. İleri Doğru Seçim'e ilişkin sonuçlar

Çizelge 3.6. Geriye Doğru Eleyerek Seçim'e ilişkin sonuçlar

Çizelge 3.7. Adımsal Seçim'e ilişkin sonuçlar

Çizelge 3.8. GA Tabanlı Seçim'e ilişkin sonuçlar

Çizelge 3.9. Lasso Regresyon Tabanlı Seçim'e ilişkin sonuçlar

Çizelge 3.10. Ridge Regresyon Tabanlı Seçim'e ilişkin sonuçlar

Çizelge 3.11. En çok kullanılan 15 değişken

Çizelge 3.12. Sağlam Değişken Seçim Yöntemi'ne ilişkin sonuçlar

ŞEKİLLER

Şekil 3.1. Değişkenler arasındaki ilişkiyi gösteren ısı haritası

Şekil 3.2. Evde kullanılan cihazların enerji kullanım miktarı değişkeninin yoğunluk grafiği

Şekil 3.3. Evde kullanılan cihazların enerji kullanım miktarı değişkeninin aylara göre keman grafiği

Şekil 3.4. Aylık ortalama enerji tüketimine ilişkin çizgi grafiği

Şekil 3.5. Chievres Havaalanı hava istasyonundan alınan dış sıcaklık grafiği (°F)

Şekil 3.6. Saatlik ortalama enerji tüketimi için çizgi grafiği

Şekil 3.7. Ortalama enerji tüketimi için ısı haritası

Şekil 3.8. Korelasyon Tabanlı Seçim sonucunda seçilen değişkenlerin saçılım grafiği matrisi

Şekil 3.9. Varyans Tabanlı Seçim sonucunda seçilen değişkenlerin saçılım grafiği matrisi

Şekil 3.10. Değişkenlerin modele olan etkisi (MSE cinsinden)

Şekil 3.11. Değişkenlerin modele olan etkisi (MSE cinsinden)

Şekil 3.12. Değişkenlerin modele olan etkisi (R^2 cinsinden)

Şekil 3.13. Oluşturulan bireylerin Nesiller boyunca MSE değerleri

Şekil 3.14. Lasso Regresyonu ile seçilen değişkenler için saçılım grafiği

Şekil 3.15. Değişkenlerin katsayılarının mutlak değerleri için çubuk grafiği

Şekil 3.16. Şekil 3.16. Katsayıların mutlak değeri için çizgi grafiği

KISALTMALAR

AI (YZ)	Artificial Intelligence (Yapay Zekâ)
AIC	Akaike Information Criterion (Akaike Bilgi Kriteri)
AICC	Corrected AIC (Düzeltilmişt AIC)
BIC	Bayesian Information Criterion (Bayesci Bilgi Kriteri)
BICC	Corrected BIC (Düzeltilmiş BIC)
CFS (KTDS)	Corellation-based Feature Selection (Korelasyon Tabanlı Değişken Seçimi)
DTs	Decision Trees (Karar Ağaçları)
GA	Genetic Algorithms (Genetik Algoritmalar)
GPT	Generative Pre-trained Transformer
K-NN	K-Nearest Neighbors (K-En Yakın Komşu)
LR (DR)	Linear Regression (Doğrusal Regresyon)
MAE (OMH)	Mean Absolute Error (Ortalama Mutlak Hata)
ML (MÖ)	Machine Learning (Makine Öğrenmesi)
MSE (HKO)	Mean Squared Error (Hata Kareler Ortalaması)
PCA (TBA)	Principal Component Analysis (Temel Bileşenler Analizi)
PCR (TBR)	Principal Component Regression (Temel Bileşenler Regresyonu)
RF (RO)	Random Forests (Rastgele Ormanlar)
SVM (DVM)	Support Vector Machines (Destek Vektör Makineleri)
SVR (DVR)	Support Vector Regression (Destek Vektör Regresyonu)

1. GİRİŞ

Günümüz teknoloji çağında, her geçen gün artan dijital aktiviteler ve teknolojik gelişmeler sayesinde üretilen ve ortaya çıkan veri miktarı adeta bir çığ gibi büyümekte ve bu durum, "büyük veri" olarak adlandırılan yeni bir çalışma alanının doğuşuna zemin hazırlamıştır. Bu büyük veri kavramı, sadece hacmiyle değil, çeşitliliği ve üretim hızıyla da geleneksel veri işleme tekniklerinin ötesine geçmiştir. Geleneksel istatistiksel yöntemler, bu verinin karmaşıklığı ve büyüklüğü karşısında yetersiz kalmakta, bu nedenle de veri bilimi disiplini içerisinde, bu devasa veri akışını etkili bir şekilde kontrol edebilmek, analiz edebilmek ve değerli bilgilere dönüştürebilmek için yeni, daha gelişmiş yöntemlerin ve teknolojilerin geliştirilmesi kaçınılmaz hale gelmiştir. Bu yeni yöntemler makine öğrenmesi ve yapay zekâ gibi alanlarda da önemli ilerlemelere yol açarak, veriden anlam çıkarma süreçlerini daha etkin ve verimli hale getirmiştir. Bu durum, veri bilimi alanının sadece akademik bir merak konusu olmaktan çıkıp, iş dünyası, sağlık, finans ve birçok diğer sektörde stratejik karar alma süreçlerinde kritik bir role sahip olmasına neden olmuştur.

Bu gelişmelerle eşanlı olarak model oluşturma süreci de çok daha karmaşık hale gelmektedir. Bu noktada modelin performansını arttırmak ve anlamlı sonuçlar elde etmek için değişken seçiminin ne kadar kritik olduğu sonucu ortaya çıkmaktadır. Yanlış değişken seçimi modelin performansını olumsuz yönde etkileyebilir ve yanıltıcı sonuçların ortaya çıkışına zemin hazırlayabilir[1].

Bunun yanı sıra, model oluşturma süreçleri de önemli ölçüde daha karmaşık bir hal almakta ve bu durum, veri bilimcileri ve analistleri için yeni zorluklar doğurmaktadır. Bu karmaşıklığın artması ve bununla birlikte modelin doğruluğunun ve performansının en iyi hale getirilmesi sorunlarına bir çözüm olarak değişken seçiminin önemi daha da artmaktadır[2]. Değişken seçimi, büyük veri kümelerinden elde edilen anlamlı ve doğru sonuçlar için kritik bir adımdır. Yanlış veya önemsiz değişkenlerin seçimi, yanıltıcı sonuçlara yol açabilir ve yanlış kararların alınmasına sebep olabilir. Bu nedenle, veri bilimi pratiklerinde, doğru değişkenleri belirleyebilmek için gelişmiş seçim teknikleri ve algoritmalarının kullanımı hayati öneme sahiptir. Bu teknikler, modelin karmaşıklığını yönetmeye, aşırı uyuma (overfitting) karşı

korumaya ve en önemlisi, performansını iyileştirmeye yardımcı olur. Özellikle, makine öğrenmesi ve yapay zekâ modellerinde, doğru değişken seçimi, modelin gerçek dünya verileri üzerindeki genelleme kapasitesini artırarak, daha güvenilir ve doğruluk oranı yüksek sonuçlar üretmesine olanak tanıyabilir. Bu süreç, veri bilimi alanında, sadece teknik becerilerin ötesinde, veriye hakimiyet ve iş süreçlerine derinlemesine anlayış gerektirir. Böylece analitik modellerden elde edilen bilgilerin iş stratejilerine ve karar alma mekanizmalarına doğrudan etkisi artırılabilir[3].

Sethi ve Mittal, Hava kalitesini etkileyen verileri incelerken Nedensellik Temelli Doğrusal (CBL) model kullanılarak değişken seçim yöntemlerini uygulamışlardır[4]. Hem tüm değişkenleri içeren hem de CBL yöntemiyle seçilen değişkenlerin bir alt kümesi ile oluşturulan model sınanmış ve bunun sonucunda model performansının iyileştiği gözlenmiştir. Tüm ML algoritmaları arasında Rastgele Ormanlar algoritması diğer algoritmalara kıyasla bağımlı değişkeni tahmin etmede en yüksek doğruluk oranına ulaşmıştır.

Kredi kartı dolandırıcılığı analizi için yapılan bir çalışmada, bilgi kazanımı (IG) ölçütünün kullanımı önerilmiştir[5]. Bu ölçüte göre bağımsız değişkeni açıklamada en başarılı olan değişkenler, Genetik Algoritma (GA) ile ML sürecine sokularak test edilir. Önerilen bu yaklaşım hassasiyet ve özgüllük açısından diğer temel teknikler ve literatürdeki güncel yöntemlerden daha iyi bir performans vermiştir.

Gazeloğlu, 2020'de yaptığı çalışmada değişken seçim yöntemlerinin kullanılmasında başarı elde etmiş bir diğer çalışmada da kalp hastalığı verileri ile sınıflandırma yapan Destek Vektör Makineleri (SVM) kullanmıştır[6]. Korelasyon Tabanlı Değişken Seçimi (CFS) uygulandığında en başarılı algoritma %84,8 oranı ile Naive Bayes olarak tespit edilmiştir. Ki-kare tabanlı özellik seçimi uygulandığında ise en başarılı algoritma %81,2 oranı ile Radial Temel Fonksiyon Ağı (RBF Network) algoritması olmuştur.

Simülasyon ile elde edilen her biri 10 değişken içeren farklı iki sınıfa ait 1000 gözlemli veri kümesinde yapılan bir diğer çalışmada, başlangıçta her değişkene eşit ağırlık atanmış ve

değişken seçim yöntemleri uygulanmasının ardından önerilen değişken seçim yönteminin önemli değişkenleri seçmede başarılı olmuştur[7].

Parkinson hastalığı analizi için kullanılan veri kümesiyle yapılan bir diğer çalışma ise 46 değişkenden oluşmaktadır[8]. Sarmalama Değişken Seçim yöntemleri kullanılmasının ardından hastalık tahmini için K-En Yakın Komşu (K-NN) algoritması kullanıldığında %88.3'lük bir doğruluk elde edildiği görülmüştür.

Chen vd. yaptıkları çalışmada 3 farklı veri kümesine değişken seçim yöntemlerinin uygulanması sonucunda Rastgele Orman (RF) tabanlı yöntem özellikle sınıflandırma doğruluğunu iyileştirmede etkili olduğunu göstermişlerdir[9].

Yapılan çalışmalardan da görüldüğü üzere değişken seçim yöntemleri her veri kümesinde aynı performansı vermemektedir. Değişkenler seçildikten sonra uygulanan tahmin ya da sınıflama algoritmalarına göre sonuç performansları farklılık göstermektedir. Bu anlamıyla veri kümesine uygulanan değişken seçim yöntemi ve sonrasında kullanılan algoritma her defasında değişmekte ve özgünlük göstermektedir. Buna bağlı olarak farklı veri kümelerinde farklı değişken seçim yöntemlerinin iyi sonuçlar verdiği çalışmalar da mevcuttur. Her veri kümesi için en iyi performansı sunan tek bir yaklaşımın olmadığı, bu nedenle farklı veri tipleri ve karakteristikleri için çeşitli yöntemlerinin denenmesi önemlidir[10].

Doğrusal regresyon analizinde sıklıkla karşılaşılan bir sorun, değişken seçimidir. Değişken seçim işlemleri, Adımsal Yöntemler (İleriye Doğru Seçim, Geriye Doğru Eleyerek Seçim, Adımsal Seçim gibi) veya tüm olası altkümeler (Mallows' Cp, AICC gibi) üzerinden gerçekleştirilebilir. Geleneksel değişken seçim yöntemleri, aykırı değerlere ve hata dağılımının normallikten sapmalarına karşı duyarlıdır. Bu nedenle, alternatif olarak Sağlam Değişken Seçim Yöntemleri önerilmektedir. Bu yöntemler, aykırı değerlerin etkisini azaltarak daha güvenilir ve doğru tahminler elde edilmesini sağlamaktadır[11][12].

Sağlam Cp, Sağlam AIC ve Wald testine dayalı seçim, Sağlam Değişken Seçim Yöntemleri'nden bazılarıdır.

Bu tez çalışması, değişken seçim yöntemleriyle değişkenlerin seçilmesi sonucunda ML algoritmalarının seçilen değişkenler üzerinde nasıl performans gösterdiğinin kapsamlı bir şekilde incelenmesine odaklanmıştır. Bu süreç, ML algoritmalarının avantajlarını, sınırlılıklarını ve uygulanabilirliklerini derinlemesine değerlendirerek, hangi senaryolar altında bir yöntemin diğerine üstünlük sağlayabileceğini belirlemeyi amaçlamaktadır. Çalışma, algoritmaların tahmin doğruluğu, işleme hızı, karmaşıklık ve genelleştirme yeteneği gibi kritik metrikler üzerinden karşılaştırmalı bir analiz sunacaktır.

Bu çalışma, ML ve istatistiksel yöntemlerin birlikte değerlendirilmesini, karşılaştırmalı ve etkili çözümlerin ortaya çıkmasına olanak tanımayı ve veri bilimi alanındaki bilgi birikimini artırırken metodolojilerin geliştirilmesine katkıda bulunmayı amaçlamaktadır.

Tez çalışmasında değişken seçim yöntemlerinden Korelasyon Analizi Tabanlı Seçim, Varyans Tabanlı Seçim, İleriye Doğru Seçim, Geriye Doğru Eleyerek Seçim, Adımsal Seçim, Genetik Algoritmalar Tabanlı Seçim, Lasso Regresyon Tabanlı Seçim, Ridge Regresyonu Tabanlı Seçim ve Sağlam M-Kestiricisine dayalı AIC, AICC, BIC ve BICC Seçim Yöntemleri kullanılmıştır. Uygulanan her değişken seçimi yöntemi sonrası farklı değişkenler seçilmiştir. Seçilen değişkenler ile ML algoritmalarından Doğrusal Regresyon, Karar Ağaçları, Rastgele Ormanlar, Destek Vektör Makineleri, Temel Bileşenler Analizi ve Yapay Sinir Ağları kullanılarak model oluşturulmuştur. Kurulan her model sonrasında bağımsız olarak eğitilmiş ve tahmin performansları tablolar halinde her bölüm sonunda gösterilmiştir. Modelin performansının ölçülmesinde Hata Kareler Ortalaması, Ortalama Mutlak Hata ve R^2 ölçütleri kullanılmıştır.

Hangi değişken seçim yöntemi ile hangi değişkenler seçilmiş olduğu; seçilen değişkenlerle kurulan modellerin ne kadar başarılı olduğu tartışılmıştır.

2. GENEL BİLGİLER

2.1. Makine Öğrenmesi

ML algoritmaları, kompleks veri kümeleri üzerinde modelleme yaparak öğrenme ve tahmin etme kabiliyetine sahip matematiksel yapılar olarak tanımlanır. Bu algoritmalar, büyük ve çeşitli veri kümelerinden karmaşık desenleri ve ilişkileri ayıklama yeteneğine sahip olup, özellikle regresyon, sınıflandırma, kümeleme ve boyut indirgeme gibi çeşitli veri işleme problemlerinde kullanılır. Regresyon analizi, sürekli değişkenler arasındaki ilişkileri tahmin ederken; sınıflandırma, verileri kategorilere ayırır. Kümeleme, benzer özellikleri olan ögeleri gruplandırır. Boyut indirgeme, veri kümelerini daha yönetilebilir ve anlamlı hale getirmek için gereksiz bilgileri filtreler. Bu süreç, veri kümelerinin daha etkin bir şekilde işlenmesini ve analiz edilmesini sağlar.

ML, temel olarak istatistik bilimine dayanır ve bilgisayarların veri aracılığıyla eğitilerek tahmin yetenekleri kazandırılmasını hedefler. Makine öğrenmesinin kökenleri, II. Dünya Savaşı dönemine ve hatta daha öncesine dayanır. Söz konusu dönemde bilgisayar bilimi ve yapay zekâ konseptleri ilk kez şekillenmeye başlamıştır. Savaş yıllarında ve sonrasında, araştırmacılar ve bilim insanları, bilgisayarların insan zekâsını taklit edebileceği ve bağımsız olarak öğrenme yeteneği gösterebileceği fikrini keşfetmeye yönelik önemli adımlar atmışlardır. Bu tarihsel arka plan, makine öğrenmesinin bugünkü gelişimine zemin hazırlayan temel teorilerin ve algoritmaların geliştirilmesinde kritik bir rol oynamıştır.

Günümüzde makine öğrenmesi, teknolojinin hemen hemen her alanında uygulanan ve sürekli olarak evrilen bir disiplin haline gelmiştir ve sağlık bilimlerinden finans sektörüne, otomotivden bilişim ve ağ alanına kadar her alanda ve geniş bir yelpazede yenilikçi çözümler sunar[13][14][15][16].

Alan Turing'in 1940'ların sonunda yazdığı makale [17], makine öğrenmesi ve yapay zekânın temellerini atmıştır. Bu dönemde, Turing bilgisayarların karmaşık sorunları çözebileceği fikrini öne sürmüş, bu da makine öğrenmesinin kökenlerini oluşturmuştur. Söz konusu makale,

bilgisayarların elektronik devreler aracılığıyla öğrenme yeteneğine sahip olabileceğini savunmuş ve yapay zekanın temel taşlarını oluşturmuştur. Turing'in çalışması, bilgisayarların sadece hesaplama yapmanın ötesinde, öğrenme ve adaptasyon yeteneklerine sahip olabileceği fikrini güçlendirmiş, bu da günümüzdeki makine öğrenmesi ve yapay zekâ uygulamalarının temelini oluşturmuştur.

Dolayısıyla bir makinenin öğrenme kapasitesi ve bunun bir sonucu olarak ortaya çıkan analiz yeteneği fikirsel ve pratik olarak hayata geçirilmiş ve doğrulanmıştır.

1950'lerde, teknolojik ilerlemeler ve işlemci kapasitelerindeki artış, makine öğrenmesi ve yapay zekâ alanlarında önemli bir ilgi artışına yol açmıştır. Bu dönemde, Arthur Samuel'in satranç oyuncuları için geliştirdiği karar ağaçları algoritması, makine öğrenmesinin temel kavramlarını somut bir şekilde ortaya koymuştur. Samuel'in çalışması [18], makinelerin kendi deneyimlerinden öğrenerek performanslarını geliştirebileceğini göstermiş ve yapay zekâ araştırmalarında yeni bir sayfa açmıştır. Bu gelişmeler, bilgisayarların sadece belirli algoritmaları izlemekle kalmayıp, aynı zamanda veri üzerinden öğrenme ve karar verme yeteneklerine sahip olabileceği fikrini pekiştirmiştir. Dolayısıyla, makine öğrenmesi ve yapay zekâ alanlarındaki bu ilk adımlar, günümüzdeki gelişmiş sistemlerin temelini oluşturmuştur.

Bu dönemde, Türk matematikçi Cahit Arf [19] makine öğrenmesinin önemine dikkat çekmiştir. Arf, bu alanın hem teknik hem de teorik yönlerine değinmiş, böylece makine öğrenmesinin sadece pratik uygulamalarla sınırlı olmadığını, aynı zamanda derin teorik temellere de sahip olduğunu vurgulamıştır. Böylece Türkiye'de makine öğrenmesi ve yapay zekâ konularına olan ilginin artmasına katkıda bulunmuş ve bu alanlarda yapılan çalışmalar için teorik bir çerçeve sunmuştur.

1970'ler ve 1980'ler, makine öğrenmesi araştırmalarında bir dönüm noktası olmuştur, bu dönemde algoritmik temeller sağlamlaştırılmış ve ileri istatistiksel yöntemlerle alan genişletilmiştir. Özellikle Gerald DeJong ve Raymond Mooney'in açıklayıcı öğrenme yöntemleri üzerine yaptıkları çalışmalar [20], makine öğrenmesi literatüründe önemli bir yer

tutmaktadır. Karar Ağaçları ve K-En Yakın Komşu (K-NN) gibi algoritmalar makine öğrenmesi araştırmalarında merkezi bir rol oynamaya başlamıştır. Karar ağaçları, veri sınıflandırma ve regresyon problemlerinde kullanılırken, K-NN algoritması, sınıflandırma ve regresyon görevlerinde basit ama etkili bir yöntem olarak kabul edilir.

Aynı zamanda, John Holland'ın öncülük ettiği genetik algoritmalar [21] gibi adapte olabilen ve kendini geliştirebilen sistemlerin araştırılması da ön plana çıkmıştır. Genetik algoritmalar, doğal seçilim ve genetik mekanizmaları taklit ederek, çözüm arama ve optimizasyon problemlerinde kullanılmak üzere tasarlanmıştır. Bu yöntemler, makine öğrenmesi ve yapay zekâ uygulamalarında, özellikle karmaşık problemlerin çözümünde ve optimizasyon görevlerinde geniş bir kullanım alanı bulmuştur.

Bu gelişmeler, makine öğrenmesinin bugünkü uygulamalarının temel taşları olmuştur. Algoritmik inovasyonlar ve istatistiksel yöntemlerin geliştirilmesi, makine öğrenmesinin daha geniş bir problem yelpazesinde etkili çözümler sunmasını sağlamıştır. Bu dönemdeki araştırmalar, makine öğrenmesinin potansiyelini genişletmiş ve gelecekteki yenilikler için zemin hazırlamıştır.

80'li yılların sonları ve 90'lı yıllar, makine öğrenmesi alanında önemli bir dönüşüm yaşanmış, özellikle Destek Vektör Makineleri (SVM) ve sinir ağları gibi algoritmaların kullanımı ve popülerliği artmıştır. Rumelhart ve arkadaşlarının sinir ağları üzerine yaptıkları çalışmalar [22], geri yayılım (backpropagation) algoritmaları kullanılarak karmaşık desenleri tanıma ve analiz etme kapasitesinin geliştirilmesi açısından dönüm noktası olmuştur. Bu, makine öğrenmesinde bir devrim yaratmış ve sinir ağlarının popüler bir araç haline gelmesini sağlamıştır.

Aynı zamanda, Cornes ve Vapnik'in geliştirdiği Destek Vektör Makineleri (SVM) [23], sınıflandırma ve regresyon problemlerinde yüksek doğruluk oranları sunarak dikkat çekmiştir. SVM, veri kümelerini en iyi şekilde ayıran bir karar sınırı (hiperdüzlem) bulmayı amaçlar ve bu özelliğiyle, özellikle yüksek boyutlu veri kümelerinde etkilidir. Bu dönemdeki gelişmeler,

makine öğrenmesi algoritmalarının, veri bilimi, görüntü işleme, doğal dil işleme gibi çeşitli alanlarda karmaşık problemleri çözmede nasıl kullanılabileceğini göstermiştir.

2000'lerin başından itibaren, büyük veri kümelerinin büyümesiyle birlikte, derin öğrenme teknolojisi öne çıkmaya başlamıştır. AlexNet [24] derin sinir ağlarının görüntü tanıma gibi görevlerde ne kadar etkili olabileceğini göstererek, görüntü ve ses tanıma ile doğal dil işleme alanlarında bir devrim başlatmıştır. Bu başarı, derin öğrenme modellerinin karmaşık problemleri çözmedeki potansiyelini gözler önüne sermiş ve akademik çevrelerle endüstriyel uygulamalar arasında derin öğrenmeye olan ilgiyi artırmıştır.

Günümüzde, makine öğrenmesi modelleri, GPT (Ön-Eğitimli Üretken Dönüştürücü-Generative Pre-trained Transformer) serisi gibi gelişmiş algoritma tekniklerini kullanarak, büyük veri kümeleri üzerinde daha doğru ve etkili sonuçlar üretebilmektedir. GPT serisi, doğal dil işleme alanında önemli bir devrim yapmış ve metin tabanlı uygulamalarda yeni standartlar belirlemiştir. Bu ilerlemeler, makine öğrenmesinin sadece teorik bir çerçeveden çıkıp, gerçek dünya uygulamalarında da etkili çözümler sunabileceğini kanıtlamıştır.

Veri miktarındaki artış, aynı zamanda değişken sayısında da önemli bir artışa yol açmıştır. Ancak, tüm değişkenlerin bağımlı değişkeni açıklama kapasitesi eşit değildir; bazıları bağımlı değişkeni çok az açıklar ya da hiç açıklamamaktadır. Bu nedenle, tahmin modeli kurulurken doğru değişkenlerin seçilmesi, modelin performansı için kritik bir öneme sahiptir. Makine öğrenmesi uygulamalarında, modelin doğruluğunu ve genel performansını artırmak için en alakalı değişkenlerin belirlenmesi ve kullanılması esastır. Bu süreç, veri bilimcileri ve araştırmacılar için model geliştirme aşamasında kilit bir aşama olup, başarılı bir makine öğrenmesi uygulamasının temelini oluşturur.

2.2. Makine Öğrenmesi Süreci

ML süreci, bir problemin tanımlanmasından modelin dağıtımına ve sürekli iyileştirilmesine kadar çeşitli adımları içerir. Bu süreç, veri toplamadan veri ön işlemeye, model eğitiminden değerlendirme ve optimizasyona kadar uzanır. Her adım, belirli görevleri ve hedefleri içerir ve bu sürecin başarısı, bu adımların dikkatli bir şekilde yürütülmesine bağlıdır[25].

Süreç ne kadar titizlikle yürütülürse model performansının başarılı olma şansı da o kadar artacaktır. Aksi takdirde model performansı aşağıda kalacaktır. Bunun yanında model performansı başarılı sonuçlar verse de verdiği sonuçların gerçek hayata uygunluğu ve doğruluğu yanıltıcı olabilir. Şimdi bu adımları açıklayalım:

2.2.1. Problem Tanımı ve Hedef Belirleme

Başarılı bir ML projesi, açıkça tanımlanmış bir problemle başlar. Bu aşamada, çözülmesi gereken sorun net bir şekilde ifade edilir ve ML modelinin bu probleme nasıl bir çözüm sunacağına dair beklentiler belirlenir. Problemin doğası (sınıflandırma, regresyon, kümeleme vb.) ve projenin iş hedefleri bu aşamada tanımlanır[26].

2.2.2. Veri Toplama

Modelin eğitimi için gerekli olan verilerin toplanması, ML sürecinin temelini oluşturur. Veri, çeşitli kaynaklardan toplanabilir; bu, kurumsal veri tabanlarından, çevrimiçi kaynaklardan, anketlerden elde edilen veriler olabilir. Veri toplama stratejisi, analiz edilecek soruna ve kullanılacak ML algoritmasına bağlı olarak değişir[27].

2.2.3. Veri Ön İşleme

Gerçek dünya verileri genelde eksik veya yanlış girilmiş olabilir. Toplanan ham verinin modele uygun hale getirilmesi, veri ön işleme aşamasında gerçekleşir. Bu süreç, veri temizleme, eksik veri işleme, aykırı değerlerin ele alınması gibi görevleri içerir. Veri ön işleme, modelin doğruluğunu ve performansını doğrudan etkileyen kritik bir adımdır[28].

2.2.4. Veri Görselleştirme

Bu adım hem kendisinden önceki için hem de kendisinden sonraki süreçler için kullanılabilir. En başta verinin yapısını görebilmek ve dağılımını anlayabilmek, değişkenler arasındaki ilişkiyi incelemek, veri ön işleme sırasında aykırı değerleri tespit edebilmek için kullanılmaktadır. Daha sonrasında hem model kurulurken hem de model değerlendirme adımında, modelin tahmin veya sınıflandırma performansını göz önüne serebilmek adına oldukça yaygın şekilde kullanılmaktadır.

2.2.5. Değişken Seçimi

Bu adım, veriden anlamlı olmayan değişken veya değişkenlerin çıkarılmasını ve seçilmesini içerir. Doğru değişkenlerin seçilmesi, modelin performansını önemli ölçüde artırabilir. Daha detaylı olarak ileride açıklanacaktır[29].

2.2.6. Model Değerlendirme

Modelin performansı, test veri kümesi üzerinde değerlendirilir. Bu değerlendirmeye göre model performansının başarılı olup olmadığına karar verilir. Bu adımda Ortalama Mutlak Hata, Hata Kareler Ortalaması, Açıklama Oranı gibi ölçütler kullanılabilir[30].

2.2.7. Model Seçimi ve Eğitimi

Uygun ML modelinin seçilmesi ve toplanan veri üzerinde eğitilmesi, bu aşamada gerçekleşir. Model seçimi, problem tanımına ve verinin niteliğine bağlıdır. Model eğitimi sırasında, algoritma belirli bir hata oranını minimize etmeye çalışırken veriden öğrenir[30].

2.3. Makine Öğrenmesi Algoritmaları

ML, bilgisayarlara veri kümelerini kullanarak tahmin yapabilme yeteneğinin kazandırılması sürecidir. Bu süreç, veriler üzerinden çıkarımlar yaparak modeller oluşturmayı içerir ve bu modellerin temeli, istatistiksel yöntemlerle sağlamlaştırılmıştır. Veri kümeleri, modelin öğrenme ve değerlendirme aşamalarında kullanılmak üzere eğitim ve test verisi olarak ikiye bölünür. Eğitim verisi kullanılarak, makinaya veri içinde gizli olan bilgiler öğretilir ve bu bilgilerle genelleme yeteneği geliştirilir. Öğrenme sürecinin tamamlanmasının ardından, test verisiyle makinenin öğrendiği bilgileri ne kadar iyi genelleyebildiği, çeşitli performans ölçütleri kullanılarak test edilir. Bu ölçütler, makinenin gerçek dünya verileri üzerindeki başarısını anlamamızı sağlar ve modelin iyileştirilmesi için yol gösterir.

ML algoritmaları Gözetimli, Gözetimsiz, Yarı-Gözetimli ve Pekiştirmeli Algoritmalar olarak dörde ayrılmaktadır:

2.3.1. Gözetimli Öğrenme Algoritmaları

Bu tür algoritmalar etiketlenmiş veri kümeleri kullanarak eğitilir. Her veri için bir girdi ve bir çıktı mevcuttur. Gözetimli Öğrenme Algoritmaları aşağıda verilmiştir:

2.3.1.1. Doğrusal Regresyon

Doğrusal Regresyon, sürekli değişkenler için kullanılır. Doğrusal bir ilişkiyi modellemek üzere verileri en iyi şekilde temsil eden bir doğru çizer[31].

Doğrusal regresyon, bir bağımlı değişken y ile bir veya daha fazla bağımsız değişken X arasındaki ilişkiyi modellemek için kullanılır ve model aşağıdaki gibi yazılır:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 \ldots + \beta_n x_n + \epsilon \quad (2.1)$$

Eşitlik (2.1)'de, β_0 kesim noktası terimi, $\beta_1, \beta_2, \ldots, \beta_n$ her bir bağımsız değişken için katsayıları ifade eder ve ϵ hata terimidir. Amaç, gözlemlenen veriye en iyi uyan β katsayılar vektörünü bulmaktır.

2.3.1.2. Lojistik Regresyon

İki veya daha fazla sınıf arasında sınıflandırma yapmak için kullanılır. Çıktılar kategoriktir[32].

Lojistik Regresyon, bağımlı değişkeninin kategorik olduğu durumlar için kullanılır. Olayın olasılığını p ile ifade edersek model aşağıdaki gibidir:

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 \pm \cdots + \beta_n x_n \quad (2.2)$$

Eşitlik (2.2)'deki model, belirli bir bağımsız değişkenler kümesi için olayın gerçekleşme olasılığını tahmin eder.

2.3.1.3. Karar Ağaçları

Karar Ağaçları, karar verme süreçlerini taklit eden hiyerarşik bir yapıdır. Hem sınıflandırma hem de regresyon problemlerinde kullanılır[33].

Belirli bir formülü olmamakla birlikte bölünme kriterleri vardır. Bunlardan ilki Entropi'dir ve aşağıdaki gibi ifade edilmektedir:

$$H(s) = -\sum p(x) \log_2 p(x)\,,\ x \in X \tag{2.3}$$

Bir diğer bölünme kriteri Gini İndeksi'dir ve aşağıdaki gibi ifade edilmektedir:

$$G(s) = 1 - \sum p(x)^2\,,\ x \in X \tag{2.4}$$

Karar ağaçları, veri kümesini daha küçük kümeler halinde bölerek çalışır ve sonunda bir karar ağacı oluşturur. Böylece tahmin performansına ulaşılır.

2.3.1.4. Rastgele Ormanlar

Rastgele Ormanlar algoritması, birden fazla karar ağacını birleştirerek oluşturulan ve gözetimli öğrenme görevleri için kullanılan bir kümelenmiş öğrenme tekniğidir. Hem sınıflandırma hem de regresyon problemleri için kullanılabilir. Rastgele ormanlar, karar ağaçlarının aşırı uyum sorununu azaltarak daha kararlı ve doğru tahminler yapmayı amaçlar[34].

Belirli bir formülü olmamakla birlikte rastgele ormanlar, birden fazla karar ağacının tahminlerini birleştirerek çalışır ve bu sayede tek bir karar ağacının yaptığı hataları düzeltmeye yardımcı olur.

Her bir karar ağacı rastgele seçilen veri alt kümeleri ve öznitelikler kullanılarak eğitilir. Bir sınıflandırma görevi için, en çok oy alan sınıf sonucu belirler; bir regresyon modeli için, ortalama tahmin değeri hesaplanır.

Rastgele ormanların temel avantajları arasında yüksek doğruluk, aşırı uyum riskinin düşük olması ve değişkenler arasındaki ilişkileri iyi bir şekilde yakalaması bulunur.

2.3.1.5. Destek Vektör Makineleri

Destek Vektör Makineleri (DVM), gözetimli öğrenme algoritmaları arasında yer alan ve özellikle sınıflandırma ve regresyon görevleri için kullanılan güçlü bir modeldir. DVM, veri kümesindeki örnekleri birbirinden ayıran optimum marjin genişliğine sahip bir hiper-düzlem bulmaya çalışır. Bu yaklaşım, modelin yeni örnekleri tahmin ederken yüksek genelleştirme performansı göstermesini sağlar[23].

SVM, veriyi sınıflandırmak için bir hiper-düzlem kullanır. İki sınıfı ayıran optimum hiper-düzlemi bulmayı amaçlar. Aşağıdaki gibi ifade edilmektedir:

$$min_{w,b} \frac{1}{2} . \|w\|^2 \tag{2.5}$$

$$y_i(w^T . x_i + b) \geq 1 \tag{2.6}$$

Eşitlik (2.6)'da w hiper-düzlemin normal vektörü; x bir değişken vektörü ve b yan (bias) terimidir. y_i sınıf etiketleridir ve (+1, -1) arasında değer almaktadır. Amaç, iki sınıf arasındaki marjini enbüyükleyecek w ve b değerlerini bulmaktır.

2.3.1.6. Destek Vektör Regresyonu

Destek Vektör Regresyonu (SVR), makine öğrenmesinde kullanılan bir regresyon tekniğidir ve SVM algoritmasının bir uzantısıdır. SVR, değişkenler arasındaki ilişkileri modellemek için ve tahmin yapmak için kullanılır[35]. SVR,

$$min_{w,b,\xi,\xi^*} \frac{1}{2}||w||^2 + C\sum_{i=1}^{n}(\xi + \xi^*) \quad (2.7)$$

biçimindedir. Eşitlik (2.7)'de *w* ağırlık vektörü, *b* sabit (yan) terim, *C* düzenleme parametresi, ξ ve ξ^* ise slack değişkenleridir. Slack değişkenler, modelin hata tolerans aralığının (epsilon) dışına çıkan veri noktalarının ölçülmesinde kullanılır. SVR'de kullanılan doğrusal regresyon fonksiyonu,

$$f(x) = w^T x + b \quad (2.8)$$

biçiminde tanımlanır.

2.3.1.7. Yapay Sinir Ağları

Bu algoritma insan beynindeki sinir ağları ile bağlanmış bulunmakta olan nöronları taklit eden bir yapıya sahiptir. Giriş katmanı (input layer), gizli katmanlar (hidden layers) ve çıkış katmanı (output layer) olmak üzere 3 katmandan oluşmaktadır. Her katmanda bulunan nöronlar veriyi işledikten sonra elde ettiği bilgiyi, sinir ağı aracılığıyla kendisinden sonraki nörona gönderirken bir aktivasyon fonksiyonu kullanır. Bu sırada iki nöron arasındaki sinir ağında oluşan ağırlık metriği bir sonraki nörona iletilir ve bu işlem çıkış katmanına kadar devam ettirilir. Bunun sonunca ortaya çıkan değer modelin tahmin performansını göstermektedir [25] ve aşağıdaki gibi ifade edilir:

$$a_i = \sigma\left[\Sigma w_{ij}.x_j + b_i\right] \tag{2.9}$$

Eşitlik (2.9)'da σ aktivasyon fonksiyonu, w_{ij} ağırlıklar, x_j girdiler ve b_i yan terimidir.

Birçok artısının yanında çalışma prensibi ve modeli eğitim maliyeti açısından olumsuz yönleri de bulunmaktadır. Modelin gizli katmanında bulunan nöronların çalışma ve karar verme süreçleri kişiler tarafından kolay yorumlanabilir değildir. Karmaşık yapısı sonucunda da model eğitiminde uzun süreler gerektirebilir[36].

2.3.2. Gözetimsiz Öğrenme Algoritmaları

Bu algoritmalar, etiketlenmemiş veri kümeleri üzerinde çalışır ve veri kümesindeki gizli yapıları veya desenleri bulmaya çalışır. Gözetimsiz Öğrenme Algoritmaları aşağıda verilmiştir:

2.3.2.1. K-Ortalama Kümeleme

K-Ortalama, veri noktalarını *k* sayıda küme oluşturacak şekilde gruplandıran bir algoritmadır. Her küme, kümeye ait noktaların ortalamasını temsil eden bir merkez etrafında toplanır[37].

Adımlar:
1. *k* merkez rastgele seçilir.
2. Her veri noktası için, en yakın küme merkezine göre bir kümeye atanır.
3. Her kümenin merkezi, o kümedeki noktaların ortalaması olarak güncellenir.
4. Küme atamaları değişmeyene kadar adım 2 ve 3 tekrarlanır.

Amaç fonksiyonu ve varyans sırasıyla aşağıdaki gibidir:

$$J = \sum_{i=1}^{k} W(C_i) \tag{2.10}$$

$$W(C_i) = \sum_{i=1}^{k} min_{\mu_{j \epsilon C}} \left(\left\| x_j - \mu_i \right\|^2 \right) \tag{2.11}$$

Eşitlik (2.10)'da J kümeleme maliyet fonksiyonu, k toplam küme sayısı, C_i kümeyi temsili, μ_i i'nci kümenin merkezi (ortalaması) ve x_j kümeye ait verileri göstermektedir[38].

2.3.2.2. Hiyerarşik Kümeleme

Hiyerarşik kümeleme, veri noktalarını bir hiyerarşi içinde gruplandırır. Bu yöntem, aglomeratif (alttan üste) veya bölücü (üstten alta) olabilir. Aglomeratif yaklaşımda, her nokta kendi başına bir küme olarak başlar ve adım adım en yakın kümeler birleştirilir. Hiyerarşik kümeleme, veri noktaları arasındaki benzerliklere dayalı bir ağaç (dendrogram) oluşturur[39].

Adımlar:

1. Başlangıçta, her veri noktası kendi başına bir küme olarak kabul edilir.
2. En yakın iki küme birleştirilir.
3. Adım 2, tüm veri tek bir kümede toplanana kadar tekrarlanır.

Benzerlik ölçüleri (örneğin, tek bağlantı, tam bağlantı) ve uzaklık metrikleri kümelerin nasıl birleştirildiğini belirler. Uzaklık metrikleri ve ifadeleri aşağıdaki gibidir.

Öklid Uzaklığı şu şekilde ifade edilmektedir:

$$d_{\text{Öklid}} = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2} \tag{2.12}$$

Manhattan Uzaklığı şu şekilde ifade edilmektedir:

$$d_{Manhattan} = \sum_{i=1}^{n} |x_i - y_i| \tag{2.13}$$

Minkowski Uzaklığı şu şekilde ifade edilmektedir:

$$d_{Minkowski} = \{\sum_{i=1}^{n}(x_i - y_i)^p\}^{1/p} \tag{2.14}$$

2.3.2.3. Temel Bileşenler Analizi

Temel Bileşenler Analizi (PCA), Boyut indirgeme için kullanılır. Verinin en çok varyansı içeren özelliklerini belirler. TBA, veri kümesinin varyansını enbüyükleyerek daha düşük boyutlu bir uzaya indirger[40].

Adımlar:
1. Veriden verinin ortalaması çıkarılır (merkezileştirme).
2. Kovaryans matrisi hesaplanır.
3. Kovaryans matrisinin özdeğerleri ve özvektörleri hesaplanır.
4. En büyük özdeğerlere karşılık gelen özvektörler, verinin yeni temelini oluşturur.

Yeni değişkenler (temel bileşenler), orijinal verinin dönüştürülmüş versiyonudur ve birbirinden bağımsızdır.

Aşağıdaki gibi ifade edilmektedir:

$$z_i = \Phi_{11}x_{i1} + \Phi_{21}x_{i1} + \cdots + \Phi_{p1}x_{ip} \quad (2.15)$$

Eşitlik (2.15)'te *i,* gözlem sayısını göstermektedir ve 1 ile n arasında değer alır. Temel bileşenler analizinde z_{i1}, z_{i2},…, z_{in} dizisinin varyansının enbüyüklenmesi amaçlanmaktadır.

$$max_{\emptyset_{11},\ldots,\emptyset_{p1}} \left\{ \frac{1}{n} \sum_{i=1}^{n} \left(\sum_{j=1}^{p} \Phi_{j1}x_{ij} \right)^2 \right\} \quad (2.16)$$

Ve katsayıların kareleri toplamı 1 olmalıdır:

$$\sum_{j=1}^{p} \emptyset_{j1}^2 = 1 \quad (2.17)$$

Temel Bileşenler Analizi'nde boyut azaltılırken varyansın anlamlı derecede yüksek olduğu boyutlar dikkate alınır. Varyansın çok düşük olduğu boyutlar ise göz ardı edilir[38].

2.3.2.4. Temel Bileşenler Regresyonu

Temel Bileşenler Regresyonu (PCR), hem çoklu doğrusal regresyon hem de PCA yöntemlerinin birleşimidir[41]. PCR özellikle yüksek boyutlu veri kümelerinde bağımsız değişkenler arasında yüksek derecede çoklu doğrusal bağlantı sorunu tespit edildiğinde kullanılır. Bu yöntem bağımsız değişkenlerin boyutunu azaltarak modelin karmaşıklığını azaltabilir. PCR,

$$Y = T\beta + \varepsilon \tag{2.18}$$

biçimindedir. Eşitlik (2.18)'de Y bağımlı değişkeni; T temel bileşenler skorlarını; β regresyon katsayı vektörlerini ve ε hata terimini temsil eder.

2.3.2. Yarı-Gözetimli Öğrenme Algoritmaları

Bu tür algoritmalar hem etiketlenmiş hem de etiketlenmemiş verileri kullanarak model eğitimini gerçekleştiren bir makine öğrenmesi yaklaşımıdır.

2.3.2.1. Kendini Eğitme

Self-Training, yarı-gözetimli öğrenme algoritmalarının en basit formudur. Bu yöntemde, bir model öncelikle etiketli verilerle eğitilir. Daha sonra, model etiketlenmemiş veriler üzerinde tahminler yapar ve bu tahminlerden yüksek güvenilirlikle yapılanları yeni etiketli veriler olarak modelin eğitim setine ekler[42].

2.3.2.2. Eş-Eğitim

Eş-Eğitim, veri kümesinin iki farklı ve yeterince bağımsız özellik setine sahip olduğu varsayımına dayanır. İki farklı sınıflandırıcı, bu iki özellik seti üzerinde ayrı ayrı eğitilir. Daha sonra, bir sınıflandırıcının etiketlenmemiş bir örnek üzerinde yüksek güvenilirlikle yaptığı tahmin, diğer sınıflandırıcı için ek etiketli veri olarak kullanılır[43].

2.3.3. Pekiştirmeli Öğrenme Algoritmaları

Pekiştirmeli öğrenme, bir ajanın çevresiyle etkileşim içinde olduğu ve gerçekleştirdiği eylemler sonucunda elde ettiği ödüller veya cezalar aracılığıyla öğrenmeye çalıştığı bir makine öğrenmesi yaklaşımıdır. Ajan, bu öğrenme sürecinde çevresinden gelen geri bildirimlere dayanarak optimal eylem politikasını bulmayı amaçlar. Özellikle oyun teorisi, robotik, otomatik kontrol sistemleri ve optimizasyon problemleri gibi alanlarda uygulama bulmuştur[44].

2.4. Değişken Seçimi

Değişken seçimi makine öğrenmesi, veri analitiği ve yapay zekâ gibi alanlarda büyük öneme sahiptir. Genelde büyük miktarda veri kullanıldığında bu yöntemlere başvurulur. Günümüzde herhangi bir alanda üretilen veriler klasik istatistiksel yöntemlerle analiz edilemeyecek kadar çok fazla hale gelmiştir. Veri miktarının büyük oluşu değişkenler için de geçerlidir. Ancak söz konusu bu bağımsız değişkenlerin bazıları, bağımlı değişkeni ya çok az miktarda açıklamaktadır ya da hiç açıklamamaktadır. Söz konusu bu durum da ileride oluşturulacak tahmin modelinin performansını, verimliliğini ve doğruluğunu olumsuz yönde etkileyebilmektedir.

Tahmin modeli kurulurken değişken seçimi yöntemleriyle en uygun değişkenlerin seçilmesi ve önemsiz değişkenlerin tespit edilip çıkartılması gerekmektedir. Bu nedenle değişken seçimi makine öğrenmesi süreçleri için en hayati adımlardan biri olarak kabul edilmektedir. Değişken seçimi yöntemlerinin uygulanması sonucunda:

Aşırı öğrenme önlenir: Çok fazla değişken modelin karmaşık olmasına neden olabilir. Karmaşık modellerde ise verinin çok iyi öğrenilmesi gibi durumlarla karşılaşılabilir. Birebir öğrenme, tahmin modelinin performansını eğitim aşamasında çok yukarılara taşırken test aşamasında çok aşağı seviyelerde olmasına neden olabilir. Aşırı öğrenme, olumlu bir durum gibi görünmekle beraber aslında tahmin modelinin genelleme yapma yeteneğinden yoksun olduğunu gösterir. Tahmin modelinin öğrendiği veri dışında yeni bir veri ile karşılaştığı durumda buna dair bir sonuç veremeyeceği anlamına gelmektedir. Gereksiz değişkenlerin veri kümesinden çıkarılması sonucunda bu durum önlenebilmektedir.

Hesap maliyeti azaltılır: Değişken sayısının çok fazla olması modelin eğitim ve test işlemlerinin daha uzun bir sürede gerçekleştirilmesine neden olabilir. İhtiyaç duyulmayan değişkenlerin söz konusu yöntemlerle veri kümesinden çıkarılması bu süreyi kısaltmaya yardımcı olacaktır.

Daha iyi bir bilgi çıkarımı: Özellikle büyük verilerde (veri madenciliği) hangi değişkenin sonucu nasıl etkilediği bilgisini elde etmek zor olabilir. Doğru değişkenlerle istenen ve en doğru bilgiye ulaşmak da kolaylaşacaktır.

2.5. Değişken Seçim Yöntemleri

Değişken seçim yöntemleri genellikle üç ana kategoriye ayrılır. Bunlar Filtreleme Yöntemleri, Sarmalama Yöntemleri ve Gömülü Yöntemler'dir.

2.5.1. Filtreleme Yöntemleri

İstatistiksel kriterler doğrultusunda değişkenler değerlendirilir. Bunun sonucunda değişken seçimi gerçekleştirilir[45].

2.5.1.1 Korelasyon Analizi Tabanlı Seçim

Regresyon ve sınıflandırma problemlerinde kullanılır. Değişkenlerin bağımlı değişkenle aralarındaki ilişki ölçülür. Düşük korelasyona sahip olan değişkenler elenir; yüksek korelasyona sahip olan değişkenler modelde tutulur[46].

Değişken seçiminde kullanılan korelasyon,

$$r = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2(y_i-\bar{y})^2}} \qquad (2.19)$$

biçiminde hesaplanır.

2.5.1.2. Ki-Kare Testi Tabanlı Seçim

Ki-Kare Testi, kategorik değişkenlerin bağımlı değişkenle olan ilişkisini değerlendirmek için kullanılır. Ki-Kare skoru yüksek olan değişkenlerin, bağımlı değişkenle anlamlı bir ilişkiye sahip olduğu kabul edilir[47].

2.5.1.3. Varyans Tabanlı Seçim

Regresyon problemleri için kullanılır ve düşük varyanslı değişkenlerin modelden çıkarılması amaçlanır. Zira düşük varyanslı değişkenler modelin tahmin yeteneğini düşürdüğü düşünülmektedir[48].

Değişken seçiminde kullanılan varyans,

$$\mathrm{Var}(x) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 \tag{2.20}$$

biçiminde hesaplanır.

2.5.2. Sarmalama Yöntemleri

Belirli değişken kombinasyonları denenerek ve adım adım eklenerek veya çıkarılarak modelin performansı iyileştirilmeye çalışılır. En iyi performansın elde edilmesi durumunda değişken seçimi tamamlanır.

2.5.2.1. İleriye Doğru Seçim

Modelde hiçbir değişken olmadan başlanır ve her adımda modelin performansını en çok artıran değişken eklenir. Bu süreç performans artışı belirli bir eşik değerin altına düşene kadar devam eder[49].

2.5.2.2. Geriye Doğru Eleyerek Seçim

Tüm değişkenlerin dahil edildiği bir modelle başlanır ve her adımda modelin performansını en az etkileyen değişken çıkarılır[46].

2.5.2.3. Adımsal Seçim

İleri Doğru Seçim ve Geriye Doğru Eleyerek Seçim yöntemlerinin bir karışımıdır. Her adımda modele değişken eklenerek veya modelden değişken çıkarılarak devam edilir[30].

2.5.2.4. Genetik Algoritmalar Tabanlı Seçim

Doğal seçilim prensipleri taklit edilerek uygun değişken kombinasyonları bulunur. Popülasyon içindeki değişken kombinasyonları, modelin performansına göre seçilir[50].

Bu değişken kombinasyonları birey olarak adlandırılır. Kombinasyon oluşturulurken bir de Fitness fonksiyonu kullanılır. Bu fonksiyon her bir kombinasyonun (bireyin) ne kadar iyi performans gösterdiğini değerlendirir. Performans ölçütü olarak MSE kullanılabilir.

2.5.3. Gömülü Yöntemler

Gömülü yöntemler, değişken seçimi sürecini modelin eğitimiyle birleştirir. Bu yöntemler, öğrenme algoritmasının kendisi tarafından belirlenen değişkenlerin önemini kullanarak hem modelin karmaşıklığını azaltır hem de tahmin performansını artırabilir.

2.5.3.1. Lasso Regresyon Tabanlı Seçim

Lasso, regresyon katsayılarına L1 cezalandırma uygulayarak hem düzenleme (regularization) sağlar hem de önemsiz değişkenleri modelden çıkararak otomatik değişken seçimi yapar. Lasso, özellikle yüksek boyutlu veri kümelerinde etkilidir çünkü bazı katsayıları tam olarak sıfıra indirebilir[51].

Lasso Regresyon hesabı şu şekilde yapılmaktadır:

$$min_{\beta} \left\{ \frac{1}{2n} \sum_{i=1}^{n} (y_i - x_i^T \beta)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right\} \quad (2.21)$$

Eşitlik (2.21)'de y_i, bağımlı değişken, x_i bağımsız değişkenler ve β regresyon katsayıları olmakla birlikte n gözlem sayısı, p değişken sayısı ve λ ceza teriminin gücünü belirleyen parametredir.

2.5.3.2. Ridge Regresyon Tabanlı Seçim

Ridge Regresyonu, L2 cezalandırması uygulayarak katsayıların büyüklüğünü kontrol altında tutar. Bu yöntem, değişken seçimi yapmaz; ancak, çoklu doğrusal bağlantı problemi olan durumlarda modelin kararlılık ve performansını artırır[52].

Ridge Regresyon parametresi, aşağıdaki modelin enküçüklenmesi ile elde edilir:

$$min_{\beta} \left\{ \frac{1}{2n} \sum_{i=1}^{n} (y_i - x_i^T \beta)^2 + \lambda \sum_{j=1}^{p} \beta^2 \right\} \quad (2.22)$$

Eşitlik (2.22)'de y_i bağımlı değişken, x_i bağımsız değişkenler ve β regresyon katsayıları olmakla birlikte n gözlem sayısı, p değişken sayısı ve λ ceza teriminin gücünü belirleyen parametredir.

2.5.3.3. Karar Ağacı Tabanlı Seçim

Bu yöntemler, her bir ağacın oluşturulması sırasında değişkenlerin önemini değerlendirir ve bir tür gömülü değişken seçimi sağlar. Rastgele ormanlar ve gradyan artırma makineleri gibi algoritmalar hangi değişkenlerin önemli olduğunu belirler[34].

2.5.4. Sağlam (Robust) Değişken Seçim Yöntemi

Sağlam Değişken Seçim Yöntemleri'nin temel amaçlarından biri, model dağılımlarından küçük sapmalara duyarsız olan yeni istatistiksel yöntemler bulmaktır.

Klasik değişken seçim ölçütleri, klasik kestirimlere dayalı olduğundan aykırı değerlere ve normallikten sapmalara oldukça duyarlıdırlar. Yani klasik yöntemler, aykırı değerlerin varlığında veya normal dağılımdan hafif sapmalarda güvenilir parametre kestirimleri vermezler. Bu durumlarda Sağlam Değişken Seçim Yöntemleri'ne ihtiyaç duyulmaktadır. Literatürde Sağlam Değişken Seçim Yöntemleri önerilmekle birlikte bu konu biraz göz ardı edilmiştir[11]. Literatürde yer alan çalışmaların bazıları şöyledir: Ronchetti, Akaike Bilgi Kriteri'nin (AIC) sağlam bir versiyonunu önermiştir[53]. Bu çalışma, aykırı değerlerin ve normal dağılımdan sapmaların mevcut olduğu veri kümelerinde AIC'in daha güvenilir sonuç vermesini sağlamayı amaçlamaktadır.

Regresyon ve otoregresif modellerde yaygın olarak kullanılan seçim yöntemlerinin başında AIC gelmektedir. Son 10 yılda, değişken seçiminde AIC'in kullanımı üzerine yoğunlaşılmıştır. AIC ile ilgili çalışmaların çoğu, normal dağılımlı veri kümelerinde ve büyük örneklemler üzerinde gerçekleştirilmiştir. Ancak, küçük örneklemli normal doğrusal regresyon modellerinde AIC yanlı olabileceğinden, bu kriterin düzeltilmiş bir versiyonu (AICC) geliştirilmiştir.[54]

Regresyon analizi için model seçiminde sıklıkla kullanılan iki kriter, model uyumu ile karmaşıklık arasında denge kurarak en uygun modeli seçmeye yardımcı olan Akaike Bilgi Kriteri (AIC)[55] ve Bayesian Bilgi Kriteri'dir (BIC)[56]. Hem AIC hem de BIC, farklı modelleri değerlendirmek ve karşılaştırmak için niceliksel bir ölçüm sağlayarak veri bilimcilerin bilinçli kararlar almasına olanak tanır. AIC,

$$AIC = 2k - 2ln\,(L) \tag{2.23}$$

biçimindedir. Eşitlik (2.23)'te *k*, modeldeki parametre sayısı ve *L,* modelin en çok olabilirlik değeridir. BIC,

$$BIC = k\,ln(n) - 2ln\,(L) \tag{2.24}$$

biçimindedir. Eşitlik (2.24)'te *k*, modeldeki parametre sayısı; *n*, örneklem büyüklüğü ve *L,* modelin en çok olabilirlik değeridir.

Düzeltilmiş Akaike Bilgi Kriteri (AICC), AIC'in küçük örneklem boyutlu durumları için uyarlanmış halidir[57]. AICC,

$$AICC = AIC + \frac{2k(k+1)}{n-k-1} \tag{2.25}$$

biçimdedir. Eşitlik (2.25)'te *k*, modeldeki parametre sayısı ve *n*, örneklem büyüklüğüdür.

Düzeltilmiş BIC (BICC), BIC'e dayanan bir model seçim kriteridir. BICC,

$$BICC = k\,ln(n) - 2\ln(L) + f(k,n) \tag{2.26}$$

biçimindedir. Eşitlik (2.26)'da *k*, modeldeki parametre sayısı; *n*, örneklem büyüklüğü; *L,* modelin en çok olabilirlik değeri ve *f(k,n)*, modelin karmaşıklığını daha hassas şekilde cezalandıran bir düzeltme terimidir.

3. UYGULAMA

3.1. Veri Kümesi

Bu çalışmada, ev aletlerinin enerji tüketimini tahmin etmek için oluşturulmuş bir veri kümesi kullanılmıştır[58].

Appliances Energy Prediction isimli veri, açık kaynak olan ve birçok veri barındıran UC Irvine Machine Learning Depository sitesinin arşivinden alınmıştır[59].

Veri, 4 buçuk ay boyunca bir evdeki tüm odalara ve dış (kuzey) cephesine konulan sensörlerden her 10 dakikada bir sıcaklık ve nem ölçülerek kaydedilmiştir. Bunun yanı sıra hava durumuna ait bilgiler Chievres Havaalanı'nda bulunan hava istasyonundan alınmıştır.

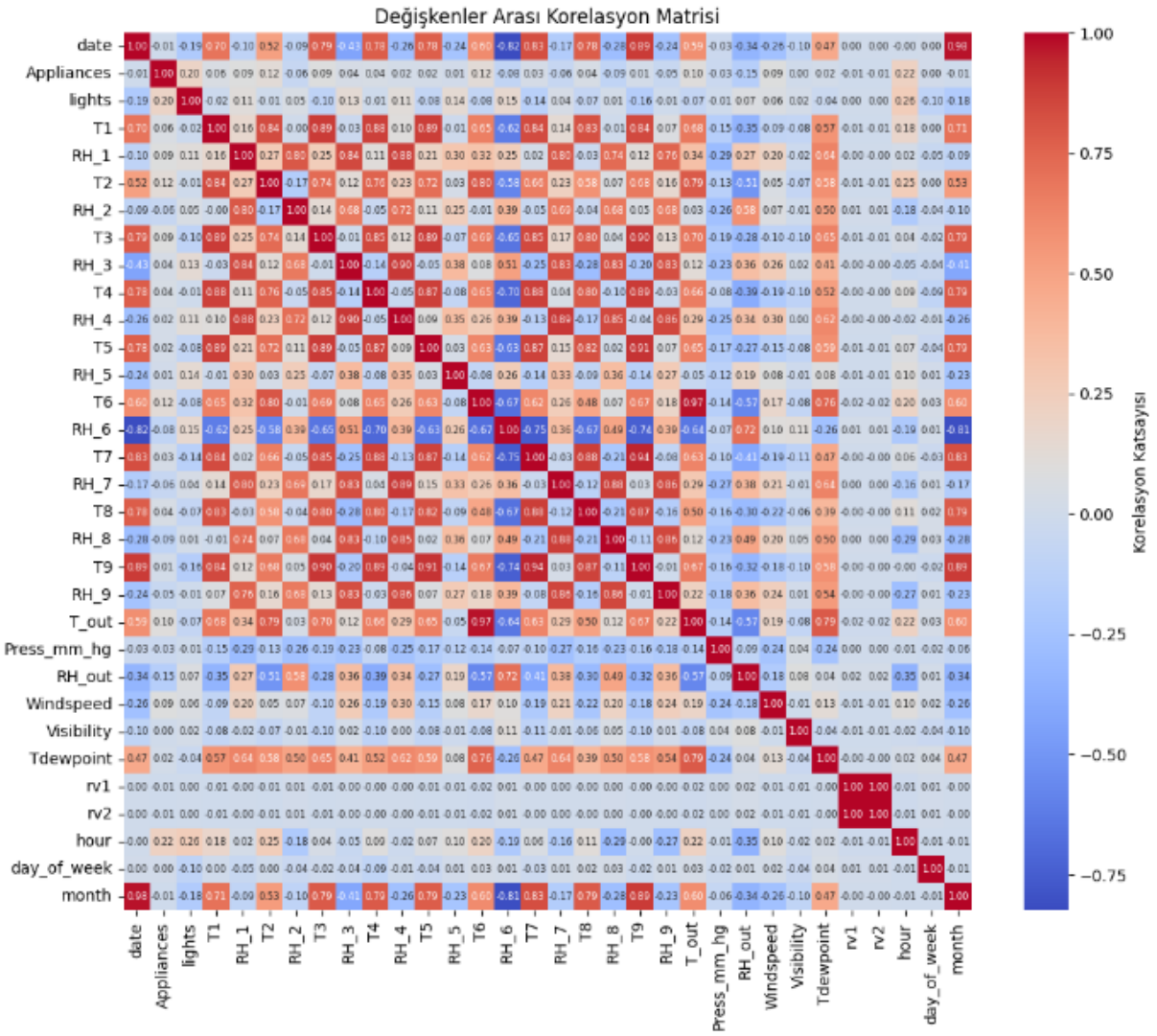
Veri kümesi 19735 gözlem ve 28 değişkenden oluşmaktadır. Veri kümesinde bulunan değişkenler Çizelge 3.1.'de gösterilmiştir

Çizelge 3.1. Değişkenlerin açıklamaları ve birimleri

Değişken İsimleri	Açıklamaları	Birim Cinsi
Date	Yıl-ay-gün-saat	
Appliances	Evde kullanılan cihazların enerji kullanım miktarı	Wh
Lights	Evde kullanılan aydınlatmanın enerji kullanım miktarı	Wh
T1	Mutfak alanındaki sıcaklık	°C
RH_1	Mutfak alanındaki nem	%
T2	Oturma odasındaki sıcaklık	°C
RH_2	Oturma odasındaki nem	%

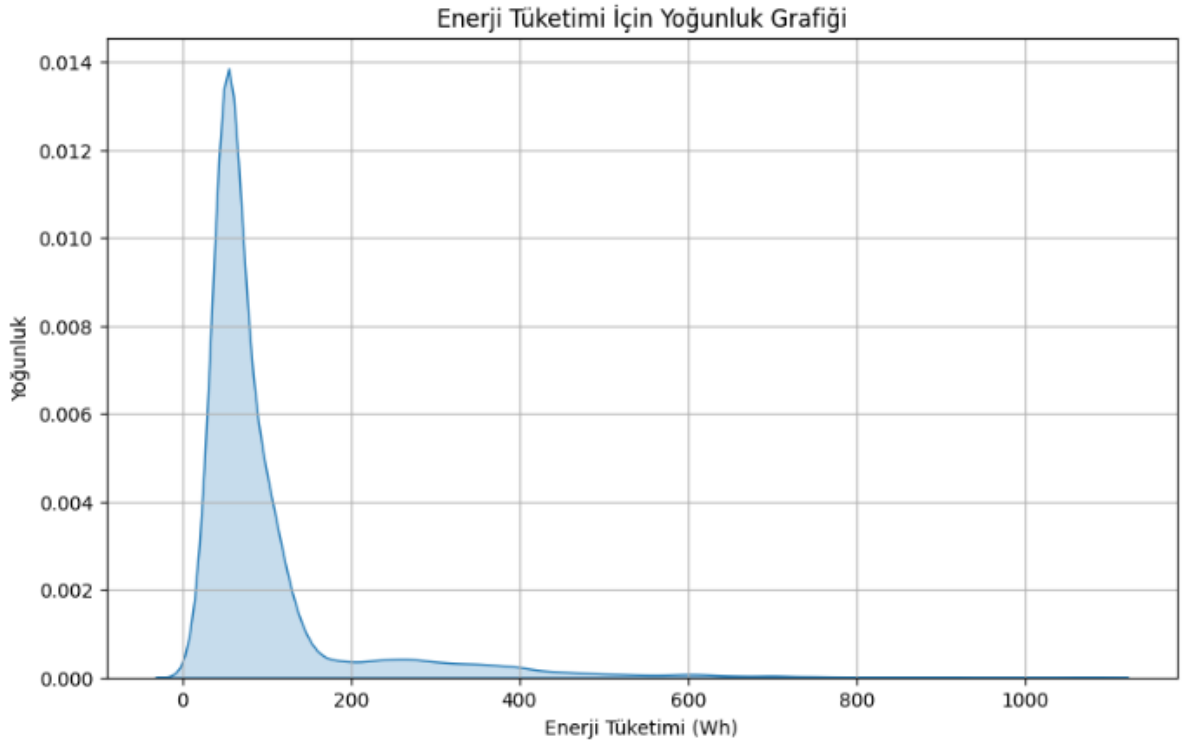
T3	Çamaşır odasındaki sıcaklık	°C
RH_3	Çamaşır odasındaki nem	%
T4	Ofis alanındaki sıcaklık	°C
RH_4	Ofis alanındaki nem	%
T5	Banyo alanındaki sıcaklık	°C
RH_5	Banyo alanındaki nem	%
T6	Binanın dışındaki (kuzey cephe) sıcaklık	°C
RH_6	Binanın dışındaki (kuzey cephe) nem	%
T7	Ütü odasındaki sıcaklık	°C
RH_7	Ütü odasındaki nem	%
T8	Genç odasındaki sıcaklık	°C
RH_8	Genç odasındaki nem	%
T9	Ebeveyn odasındaki sıcaklık	°C
RH_9	Ebeveyn odasındaki nem	%
T_Out	Dış sıcaklık (Chievres hava istasyonundan)	°C
RH_Out	Dış nem (Chievres hava istasyonundan)	%
Press_mm_hg	Basınç (Chievres hava istasyonundan)	Mm Hg
Windspeed	Rüzgâr hızı (Chievres hava istasyonundan)	M/s
Visibility	Görüş mesafesi (Chievres hava istasyonundan)	Km
Tdewpoint	Çiy düşme noktası sıcaklığı (Chievres hava istasyonundan)	°C
rv1	Rastgele değişken 1	
rv2	Rastgele değişken 2	

Şekil 3.1.'de tüm değişkenlere ilişkin ısı haritası grafiği çizdirilmiştir.



Şekil 3.1. Değişkenler arasındaki ilişkiyi gösteren ısı haritası

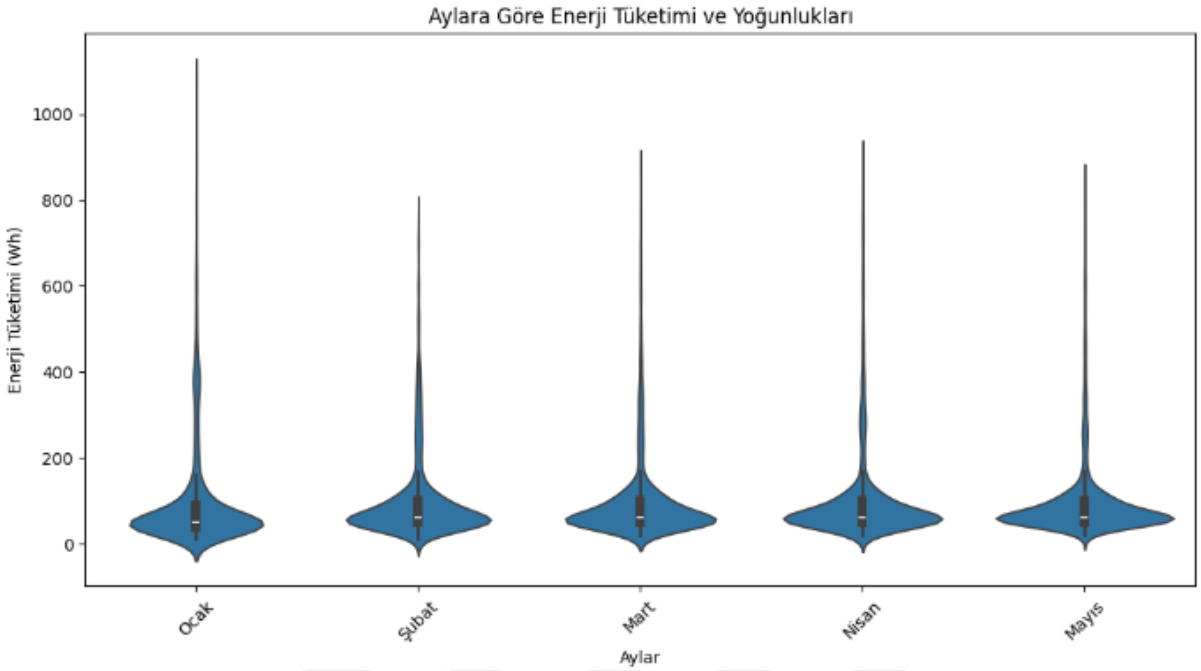
Şekil 3.2.’de Evde kullanılan cihazların enerji kullanım miktarına (bağımlı değişken) ilişkin yoğunluk grafiği verilmiştir.



Şekil 3.2. Evde kullanılan cihazların enerji kullanım miktarı değişkeninin yoğunluk grafiği

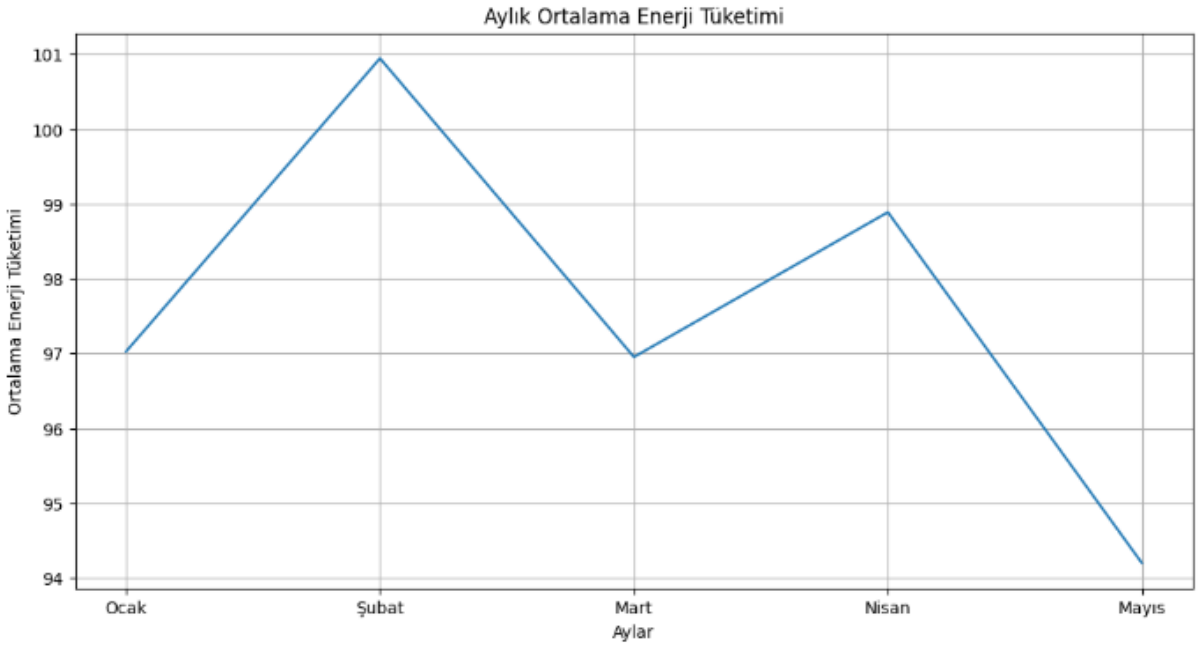
Şekil 3.2’de görüldüğü üzere bağımlı değişken normal dağılmamaktadır.

Şekil 3.3.’te evde kullanılan cihazların enerji kullanım miktarı değişkeninin aylara göre dağılımı için çizdirilen keman (violin) grafiği verilmiştir.



Şekil 3.3. Evde kullanılan cihazların enerji kullanım miktarı değişkeninin aylara göre violin grafiği

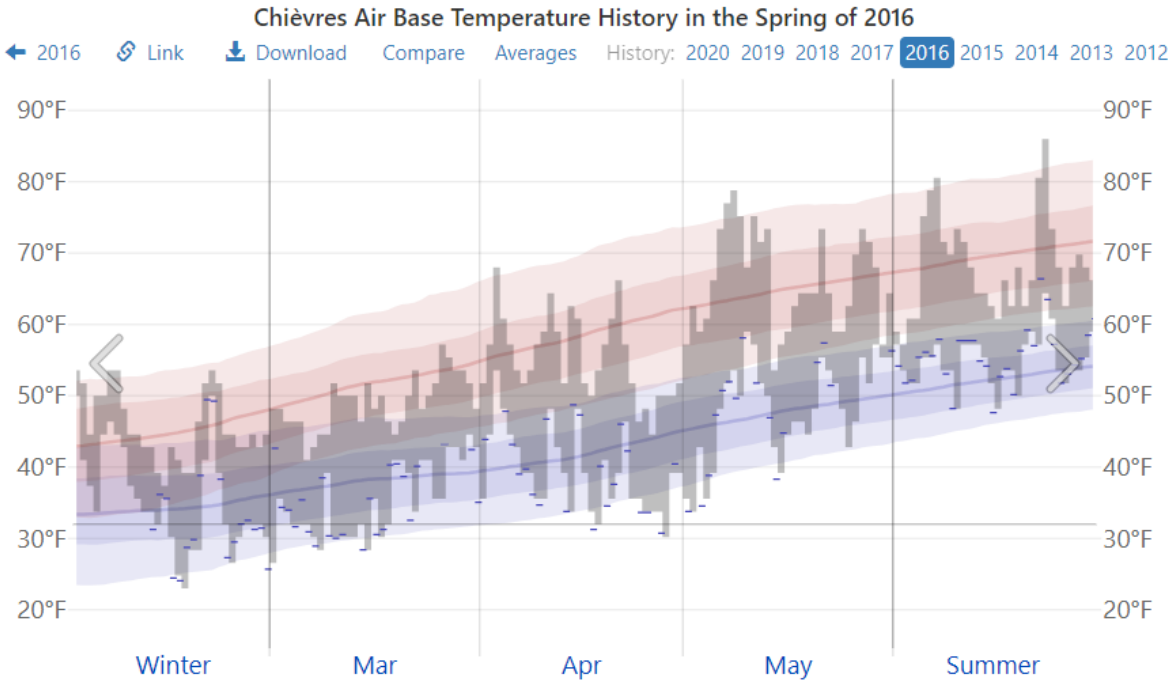
Evde kullanılan cihazların enerji kullanım miktarı değişkeninin aylara göre ortalama değerlerine ilişkin çizdirilen çizgi grafiği Şekil 3.4.'te verilmiştir.



Şekil 3.4. Aylık ortalama enerji tüketimine ilişkin çizgi grafiği

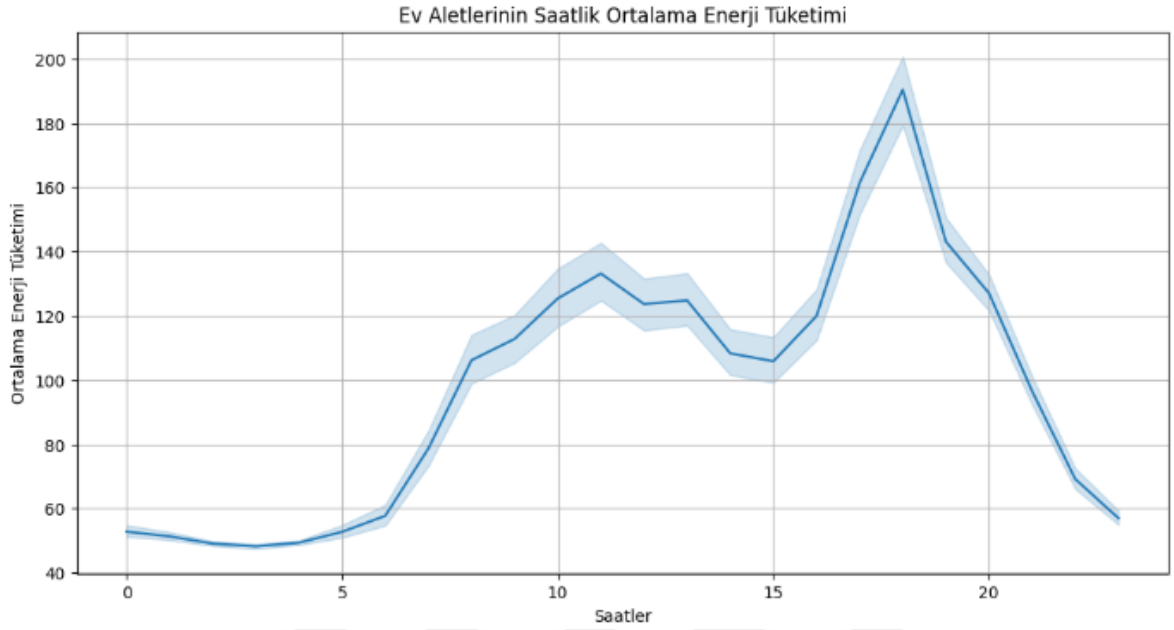
Şekil 3.4.'teki grafikten Mart Ayı'nda enerji tüketiminin düştüğü, Nisan Ayı'nda Mart Ayı'na göre yükseldiği görülmektedir. Enerji tüketiminin Ocak Ayı'ndan Mayıs Ayı'na doğru doğrusal olarak düşmesi beklendiğinden bu anormal bir durum olarak görülebilir.

Ancak bununla birlikte 2016 senesinde yapılan meteoroloji ölçümleri[60] incelendiğinde görüleceği üzere Mart Ayı'nda sıcaklıklar sıfır santigrat derecenin altında iken üstüne doğru çıkmaktadır. Nisan Ayı'nda bu durumun tersine döndüğü, sıcaklıkların tekrardan sıfır santigrat derecenin altına düştüğü görülebilir. Bu da ev içinde kullanılan ısıtıcılarla birlikte enerji tüketiminin neden Nisan Ayı'nda Mart Ayı'na göre fazla olduğunu açıklayabilmektedir. Şekil 3.5'teki grafik 2016 senesinde yapılan meteoroloji ölçümleri verilmiştir.



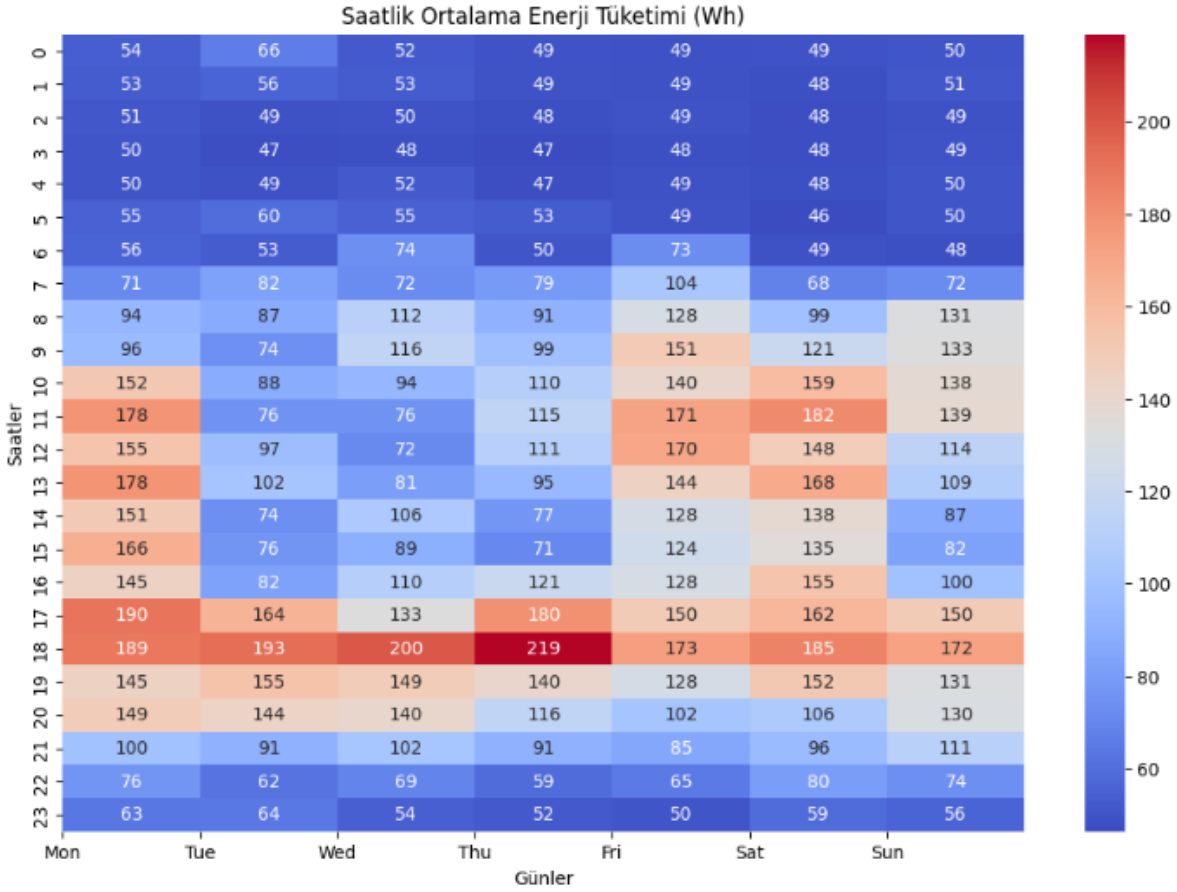
Şekil 3.5. Chievres Havaalanı hava istasyonundan alınan dış sıcaklık grafiği (°F)

Şekil 3.6.'te Evde kullanılan cihazların enerji kullanım miktarı değişkeninin saatlere göre ortalama değerlerini gösterilmesi amacıyla çizgi grafiği çizdirilmiştir.



Şekil 3.6. Saatlik ortalama enerji tüketimi için çizgi grafiği

Şekil 3.7.'de Evde kullanılan cihazların enerji kullanım miktarı değişkeninin saatlere göre ortalama değerlerinin gösterilmesi için ısı haritası çizdirilmiştir.



Şekil 3.7. Ortalama enerji tüketimi için ısı haritası

3.2. Yöntem

Bu tez çalışmasında amaç, değişken seçim yöntemlerini kullanarak ham veriden değişkenleri seçmek, model kurmak ve ardından ML algoritmalarını uygulayarak kurulan modellerin performanslarını karşılaştırmaktır.

Bu çalışmada kullanılan veri kümesine ait tanımlayıcı istatistikler, grafikler, değişken seçim yöntemleri ve ML algoritması sonucunda ortaya çıkan performans başarı miktarları Jupyter Lab ortamı kullanılarak yapılmıştır. Jupyter Lab'ın temelinde çalışan Python programının sürümü 3.12'dir.

Program içinde veri manipülasyonu ve görselleştirmeler için NumPy, Pandas, Plotly, Matplotlib ve Seaborn kütüphaneleri kullanılmıştır. Değişken seçimi ve ML algoritmalarının kurulması aşamasında Scikit-Learn, StatsModels, TensorFlow ve Keras kütüphaneleri kullanılmıştır.

Sağlam Değişken Seçimi Yöntemi uygulaması sırasında R programlama dilinin MASS ve RobustBase kütüphaneler kullanılarak değişken seçimi gerçekleştirilmiştir. Program R Studio ortamında çalıştırılmıştır ve sürümü 4.2.3'tür.

Kullanılacak yöntemler, algoritmalar ve performans ölçütleri Çizelge 3.2'de özetlenmiştir.

Çizelge 3.2. Kullanılan Yöntem, Algoritma ve Ölçütler

Değişken Seçim Yöntemi	ML Algoritmaları	Performans Ölçütleri
Korelasyon Tabanlı Varyans Tabanlı İleriye Doğru Geriye Doğru Eleyerek Adımsal Genetik Algoritmalar Tabanlı Lasso Regresyon Tabanlı Ridge Regresyon Tabanlı Sağlam Yöntem	Lineer Regresyon Karar Ağaçları Rastgele Ormanlar Destek Vektör Makineleri Temel Bileşenler Analizi Yapay Sinir Ağları	MAE MSE R^2

Çizelge 3.2 de verilen değişken seçim yöntemleriyle değişkenler seçilmiştir. Her değişken seçim yöntemi sonucunda farklı değişken kombinasyonları ortaya çıkmıştır. Bu seçim işleminden sonra ML algoritmaları uygulanmadan önce veri eğitim ve test verisi olarak ayrılmıştır. Verinin %80'i ile veri eğitilmiştir. Geri kalan %20'lik kısmıyla da test edilmiştir. Her algoritma öncesinde bu işlem gerçekleştirilmiştir. Veri eğitilip test edildikten sonra performans ölçütleri kullanılarak modelin başarısı ölçülmüştür.

Kullanılan performans ölçütleri aşağıda verilmiştir:

Ortalama Mutlak Hata (MAE):

MAE, gerçek değer ile tahmin edilen değer arasındaki farkın mutlak değerini alır. Tüm gözlemler için bu işlem yapılır ve toplanır. Son olarak gözlem sayısına bölünerek bulunur:

$$\frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{3.1}$$

Bu değerin düşük olması model performansının daha başarılı olduğunu göstermektedir.

Hata Kareler Ortalaması (MSE):

MSE, gerçek değer ile tahmin edilen değer arasındaki farkın karelerinin toplamlarının gözlem sayısına bölünmesiyle elde edilir:

$$\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \quad (3.2)$$

Bu değerinin düşük olması model performansının daha başarılı oluğunu göstermektedir.

Açıklama Oranı (R^2):

Açıklama Oranı ya da Belirtme Katsayısı, özellikle regresyon analizi modellerinde bağımlı değişkenin bağımsız değişken veya değişkenler tarafından hangi miktarda açıklandığını göstermektedir.

Açıklama Oranı,

$$1 - \frac{KT_{Hata}}{KT_{Toplam}} \quad (3.3)$$

$$1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y}_i)^2} \quad (3.4)$$

biçimindedir.

Açıklama Oranı değerinin yüksek olması model performansının daha başarılı olduğunu göstermektedir.

3.3. Makine Öğrenmesi Algoritmaları Parametreleri

3.3.1. Doğrusal Regresyon Algoritması

Scikit-Learn kütüphanesi içinde bulunan *LinearRegression* algoritması varsayılan ayarlarda kullanılmıştır. Model aşağıda belirtilen varsayımlar altında çalışmaktadır:

fit_intercept=True, Modelin sabit bir terim içerip içermeyeceğini belirler.

normalize='deprecated', Verilerin normalize edilip edilmeyeceğini belirtir.

copy_X=True, Modelin, bağımsız değişkenlerin bir kopyasını oluşturup oluşturmayacağını belirler.

n_jobs=None, Hesaplamalarda kullanılacak iş parçacığı sayısını belirler.

3.3.2. Karar Ağaçları Algoritması

Scikit-Learn kütüphanesi içinde bulunan *DecisionTreeRegressor* algoritması varsayılan ayarlarda kullanılmıştır. Model aşağıda belirtilen varsayımlar altında çalışmaktadır:

criterion='squared_error', Bölme kriteri: Hata Kareler Ortalaması.

splitter='best', Hangi bölme stratejisinin kullanılacağını belirtir: 'best' (en iyi) veya 'random' (rastgele).

max_depth=None, Ağacın maksimum derinliği.

min_samples_split=2, Bir düğümü bölmek için gereken minimum örnek sayısı.

min_samples_leaf=1, Bir yaprak düğümünde bulunması gereken minimum örnek sayısı.

min_weight_fraction_leaf=0.0, Bir yaprak düğümünde bulunması gereken minimum ağırlık (örneklerin toplam ağırlık oranı).

max_features=None, Bölme sırasında dikkate alınacak maksimum özellik sayısı.

random_state=None, Rastgele sayı üreteci durumu.

max_leaf_nodes=None, Maksimum yaprak düğümü sayısı.

min_impurity_decrease=0.0, Bir düğümün bölünmesi için gereken minimum saflık azalması.

ccp_alpha=0.0, Karmaşıklık budama parametresi.

3.3.3. Rastgele Ormanlar Algoritması

Scikit-Learn kütüphanesi içinde bulunan *RandomForestRegressor* algoritması varsayılan ayarlarda kullanılmıştır. Model aşağıda belirtilen varsayımlar altında çalışmaktadır:

n_estimators=100, Ormandaki ağaç sayısı.

criterion='squared_error', Bölme kriteri: Hata Kareler Ortalaması.

max_depth=None, Her bir ağacın maksimum derinliği.

min_samples_split=2, Bir düğümü bölmek için gereken minimum örnek sayısı.

min_samples_leaf=1, Bir yaprak düğümünde bulunması gereken minimum örnek sayısı.

min_weight_fraction_leaf=0.0, Bir yaprak düğümünde bulunması gereken minimum ağırlık (örneklerin toplam ağırlık oranı).

max_features='auto', Her bölme sırasında dikkate alınacak maksimum özellik sayısı.
max_leaf_nodes=None, Maksimum yaprak düğümü sayısı.

min_impurity_decrease=0.0, Bir düğümün bölünmesi için gereken minimum saflık azalması.

bootstrap=True, Bootstrap örneklerinin kullanılıp kullanılmayacağı.

oob_score=False, OOB (Out-of-bag) skoru kullanılıp kullanılmayacağı.

n_jobs=None, Paralel işler için kullanılacak iş parçacığı sayısı.

random_state=None, Rastgele sayı üreteci durumu.

verbose=0, Öğrenme sürecinin ayrıntılı çıktılarının olup olmayacağı.

warm_start=False, Önceki fit sonuçlarının kullanılıp kullanılmayacağı.

ccp_alpha=0.0, Karmaşıklık budama parametresi.

max_samples=None, Her ağacı eğitmek için kullanılacak maksimum örnek sayısı.

3.3.4. Destek Vektör Makineleri Algoritması

Scikit-Learn kütüphanesi içinde bulunan *SVM* algoritması varsayılan ayarlarda kullanılmıştır. Model aşağıda belirtilen varsayımlar altında çalışmaktadır:

kernel='linear', Çekirdek tipi.

gamma='scale', Çekirdek katsayısı.

coef0=0.0, Çekirdek fonksiyonu için bağımsız terim.

tol=1e-3, Çözüm doğruluğu için tolerans.

C=1.0, Düzenleme parametresi.

epsilon=0.1, Eğriye uygunluk için epsilon tüpü.

shrinking=True, Shrinking sezgisel yöntemin kullanılıp kullanılmayacağı.

cache_size=200, Çekirdek dönüşümü için bellek boyutu (MB).

verbose=False, Detaylı çıkış olup olmayacağı.

max_iter=-1, İzin verilen maksimum iterasyon sayısı (sınırsız).

3.3.5. Temel Bileşenler Analizi Algoritması

Scikit-Learn kütüphanesi içinde bulunan *PCA* algoritması varsayılan ayarlarda kullanılmıştır. Model aşağıda belirtilen varsayımlar altında çalışmaktadır:

n_components=None, Dönüştürülecek bileşen sayısı.

copy=True, X'in kopyasının oluşturulup oluşturulmayacağı.

whiten=False, Beyazlatma uygulanıp uygulanmayacağı.

svd_solver='auto', SVD çözücüsü.

tol=0.0, SVD hesaplaması için tolerans.

iterated_power='auto', Güç iterasyonu sayısı.

random_state=None, Rastgele sayı üreteci durumu.

3.3.6. Yapay Sinir Ağları Algoritması

Scikit-Learn kütüphanesi içinde bulunan *MLPRegressor* algoritması varsayılan ayarlarda kullanılmıştır. Model aşağıda belirtilen varsayımlar altında çalışmaktadır:

Dense(10, input_dim=X_train_scaled.shape[1], activation='relu'), Giriş katmanı.

Dense(5, activation='relu'), Gizli katman.

Dense(1, activation='linear', Çıkış katmanı.

activation='relu', # Aktivasyon fonksiyonu.

solver='adam', Ağırlık optimizasyon algoritması.

alpha=0.0001, L2 ceza parametresi.

batch_size=50, Mini-batch boyutu.

learning_rate='constant', Öğrenme hızı: 'constant', 'invscaling', 'adaptive'.

learning_rate_init=0.001, Başlangıç öğrenme hızı.

power_t=0.5, Öğrenme hızının invscaling sırasında düşme hızı.

max_iter=200, Maksimum iterasyon sayısı.

shuffle=True, Eğitim verilerinin her iterasyonda karıştırılıp karıştırılmayacağı.

random_state=None, Rastgele sayı üreteci durumu.

tol=1e-4, Eğitim durma kriteri.

verbose=False, Detaylı çıkış olup olmayacağı.

warm_start=False, Önceki fit sonuçlarının kullanılıp kullanılmayacağı.

momentum=0.9, Momentum değeri.

nesterovs_momentum=True, Nesterov momentumunun kullanılıp kullanılmayacağı.

early_stopping=False, Erken durdurma olup olmayacağı.

validation_fraction=0.1, Erken durdurma için doğrulama veri oranı.

beta_1=0.9, Adam çözümleyicisi için beta_1 değeri.

beta_2=0.999, Adam çözümleyicisi için beta_2 değeri.

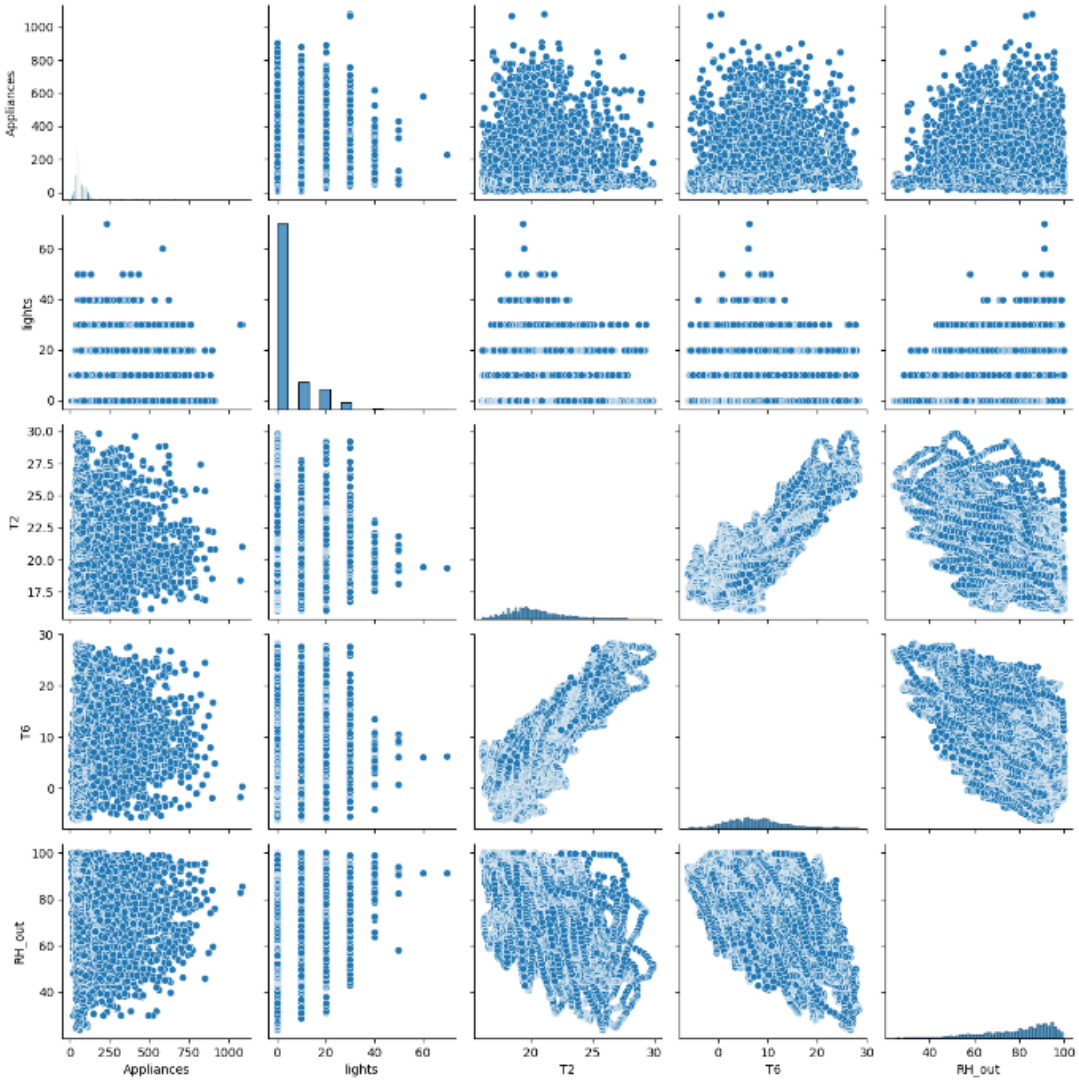
epsilon=1e-8, Adam çözümleyicisi için epsilon değeri.

3.4. Model Kurulum ve Performansları

3.4.1. Korelasyon Tabanlı Seçim

Korelasyon Tabanlı Seçim yöntemi uygulanması sonucunda bağımlı değişken ile aralarında yüksek korelasyona sahip olan **lights, T2, T6 ve RH_out** değişkenleri seçilmiştir.

Şekil 3.8.'de aralarındaki ilişkiyi gösteren saçılım grafiklerinden oluşan matris çizdirilmiştir.



Şekil 3.8. Korelasyon Tabanlı Seçim sonucunda seçilen değişkenlerin saçılım grafiği

Korelasyon Tabanlı Seçim sonucunda ML algoritmaları ile kurulan modeller ve performans değerleri Çizelge 3.3.'te verilmiştir.

Çizelge 3.3. Korelasyon Tabanlı Seçim'e ilişkin sonuçlar

Algoritma/Ölçüt	**MAE**	**MSE**	**R^2**
Doğrusal Regresyon	54.11	9065.12	0.01
Karar Ağaçları	42.66	10178.75	0.00
Rastgele Ormanlar	**36.32**	**5264.74**	**0.47**
Destek Vektör Makineleri	43.14	9707.72	0.03
Temel Bileşenler Analizi	59.37	14464.55	0.00
Yapay Sinir Ağları	52.15	8646.88	0.13

Çizelge 3.3.'te ML algoritmalarıyla kurulan modeller arasında Doğrusal Regresyon modeli çok düşük bir R^2 değerine sahiptir. Bu da modeldeki bağımsız değişkenlerin bağımlı değişkeni açıklamadaki performansının çok düşük olduğunu göstermektedir. Aynı durum Destek Vektör Makineleri ve Yapay Sinir Ağları modelleri için de geçerlidir.

Temel Bileşenler Analizi ve Karar Ağaçları modelleri için R^2 değerleri sıfırdır. Bu algoritmalar için bağımsız değişkenlerle bağımlı değişken arasında ilişki yoktur. Bu iki model için MAE ve MSE değerleri çok yüksek çıkmıştır.

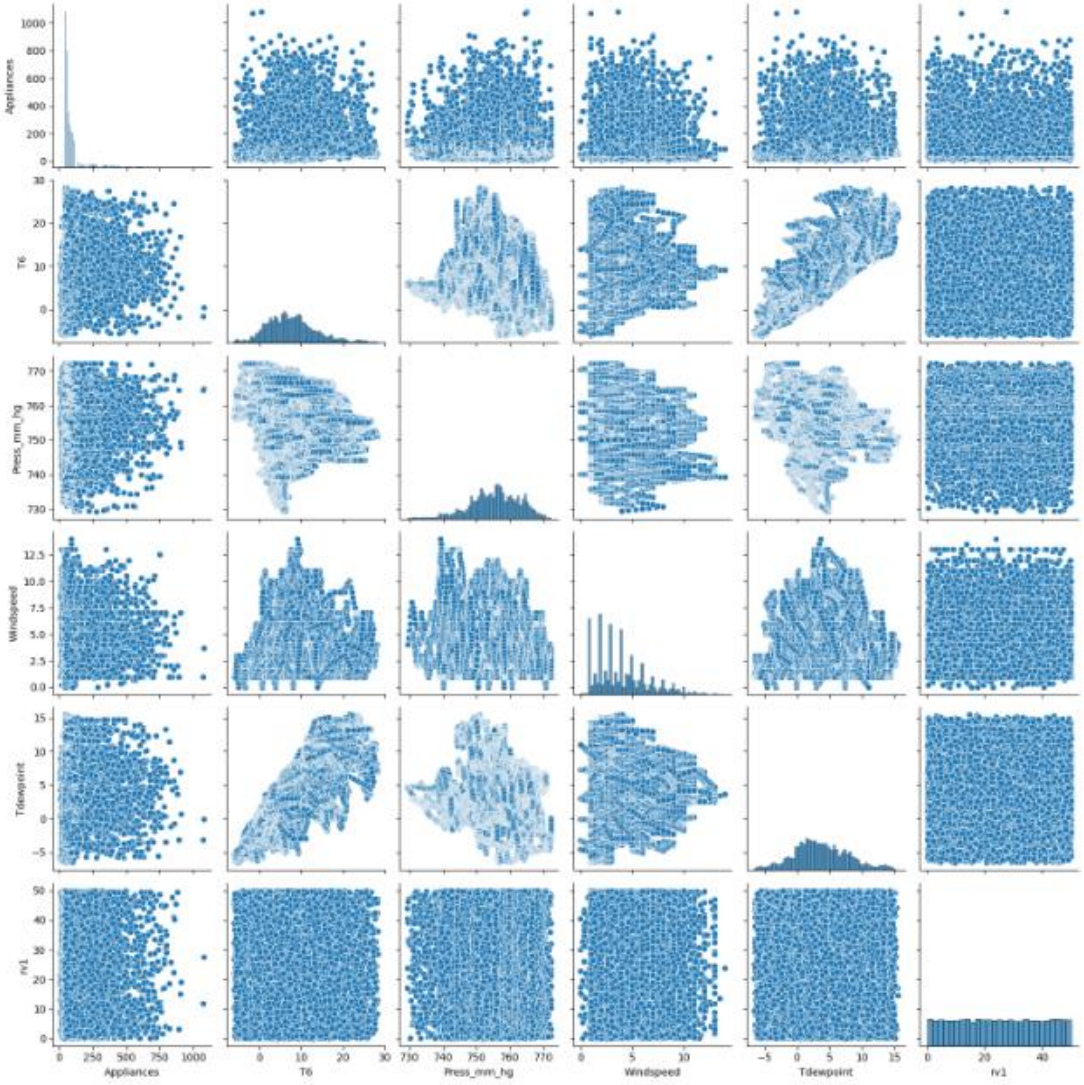
Tüm bu modellerin aksine Rastgele Ormanlar modeli diğer modellere kıyasla en iyi performansa sahip model olmuştur. Modeldeki bağımlı değişkenin, sadece 4 adet bağımsız değişken tarafından açıklanma oranı %47'dır ve en yüksek R^2 değerine sahiptir. MAE ve MSE değerleri de diğer modellere göre düşüktür.

Korelasyon Tabanlı Seçim yöntemi sonucunda seçilen değişkenlerle kurulan en başarılı model Rastgele Ormanlar Algoritması ile kurulan modeldir. Diğer algoritmalarla kurulan model performansları oldukça zayıf veya geçersiz kalmıştır.

3.4.2. Varyans Tabanlı Seçim

Varyans Tabanlı Seçim yöntemi uygulanması sonucunda düşük varyanslı değişkenler çıkarıldıktan sonra görece daha yüksek varyanslı **T6, Press_mm_hg, Windspeed, Tdewpoint ve rv1** değişkenleri seçilmiştir.

Şekil 3.9.'da seçilen değişkenlerin aralarındaki ilişkiyi gösteren saçılım grafiklerinden oluşan matris çizdirilmiştir.



Şekil 3.9. Varyans Tabanlı Seçim sonucunda seçilen değişkenlerin saçılım grafiği matrisi

Şekil 3.9'a bakıldığında bağımlı değişken ile bağımsız değişkenler arasındaki ilişkinin rastgele olduğu görülmektedir. Tdewpoint değişkeni ile T6 değişkeni arasında pozitif ve doğrusal bir ilişki vardır.

Varyans Tabanlı Seçim sonucunda ML algoritmaları ile kurulan modeller ve performans değerleri Çizelge 3.4.'te verilmiştir.

Çizelge 3.4. Varyans Tabanlı Seçim'e ilişkin sonuçlar

Algoritma/Ölçüt	**MAE**	**MSE**	**R^2**
Doğrusal Regresyon	58.79	9736.65	0.03
Karar Ağaçları	71.94	19074.84	0.00
Rastgele Ormanlar	**37.56**	**5164.75**	**0.48**
Destek Vektör Makineleri	48.05	10847.70	0.00
Temel Bileşenler Analizi	58.79	9736.80	0.03
Yapay Sinir Ağları	56.69	9902.33	0.01

Çizelge 3.4.'te ML algoritmalarıyla kurulan modeller arasında Doğrusal Regresyon modeli çok düşük bir R^2 değerine sahiptir. Bu da modeldeki bağımsız değişkenlerin bağımlı değişkeni açıklamadaki performansının çok düşük olduğunu göstermektedir. Aynı durum Temel Bileşenler Analizi ve Yapay Sinir Ağları modelleri için de geçerlidir.

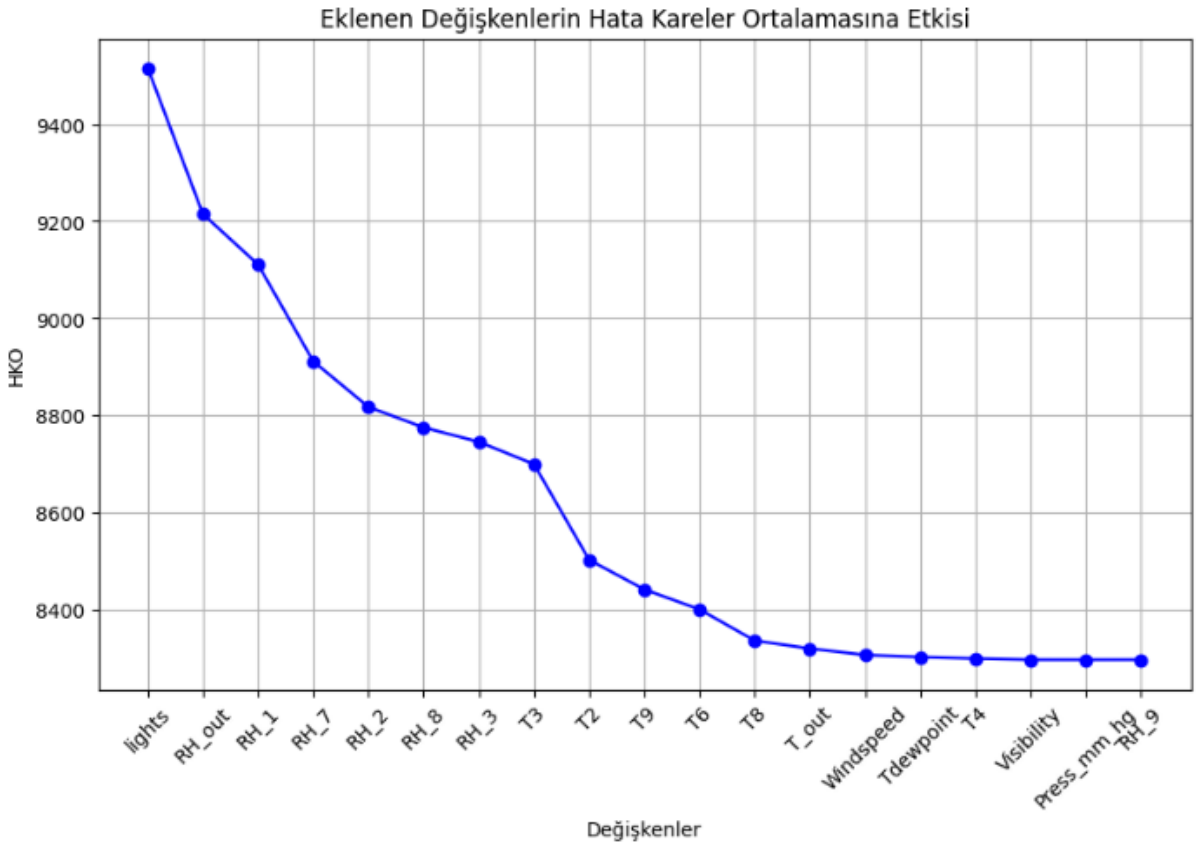
Destek Vektör Makineleri ve Karar Ağaçları modelleri için R^2 değerleri sıfırdır. Bu algoritmalar için bağımsız değişkenlerle bağımlı değişken arasında ilişki yoktur. Bu iki model için MAE ve MSE değerleri çok yüksek çıkmıştır.

Tüm bu modellerin aksine Rastgele Ormanlar modeli diğer modellere kıyasla en iyi performansa sahip model olmuştur. Bağımlı değişkenin, sadece 5 bağımsız değişken tarafından açıklanma oranı %48'dır ve en yüksek R^2 değerine sahiptir. MAE ve MSE değerleri de diğer modellere göre neredeyse yarı yarıya düşüktür.

Varyans Tabanlı Seçim yöntemi sonucunda seçilen değişkenlerle kurulan en başarılı model Rastgele Ormanlar Algoritması ile kurulan modeldir. Diğer algoritmalarla kurulan model performansları oldukça zayıf veya geçersiz kalmıştır.

3.4.3. İleriye Doğru Seçim

İleriye Doğru Seçim yönteminin uygulanması sonucunda seçilen değişkenler için modelin MSE değerini ne kadar etkilediklerini göstermek için Şekil 3.10.'daki grafik çizdirilmiştir.



Şekil 3.10. Değişkenlerin modele olan etkisi (MSE cinsinden)

Modelde hiç değişken yokken değişkenler teker teker eklenmiştir. Bu işlem sırasında modele ait MSE değeri baz alınmıştır. Değişkenler modele eklenirken MSE'yi düşüren her değişken modele katılmıştır. İleriye Doğru Seçim yöntemi sonucunda Şekil 3.10.'daki grafiğin x-ekseninde yer alan tüm değişkenleri seçmiştir. Ancak modele ait MSE değerini etkilemeyen T_out, Windspeed, Tdewpoint, T4, Visibility, Press_mm_hg ve RH_9 değişkenleri modele dahil edilmemiştir. Dolayısıyla bu işlemler sonucunda **lights, RH_out, RH_1, RH_7, RH_2, RH_8, RH_3, T3, T2, T9, T6 ve T8** değişkenleri seçilmiştir.

İleriye Doğru Seçim sonucunda ML algoritmaları ile kurulan modeller ve performans değerleri Çizelge 3.5.'te verilmiştir.

Çizelge 3.5. İleriye Doğru Seçim'e ilişkin sonuçlar

Algoritma/Ölçüt	**MAE**	**MSE**	**R^2**
Doğrusal Regresyon	52.68	8335.55	0.16
Karar Ağaçları	41.24	9887.94	0.01
Rastgele Ormanlar	**33.21**	**4921.71**	**0.51**
Destek Vektör Makineleri	43.28	9589.31	0.04
Temel Bileşenler Analizi	55.47	9094.53	0.09
Yapay Sinir Ağları	50.24	7906.81	0.21

Çizelge 3.5.'te ML algoritmalarıyla kurulan modeller arasında Doğrusal Regresyon modeli bir önceki modellere göre performans artışı göstermiş olsa da yine de düşük bir R^2 değerine sahiptir. Aynı durum Yapay Sinir Ağları ile oluşturulan model için de geçerlidir. Bu da modeldeki bağımsız değişkenlerin bağımlı değişkeni açıklamadaki performansının düşük olduğunu göstermektedir. Bu iki algoritmanın oluşturduğu modellerde MAE ve MSE değerleri de birbirine yakındır.

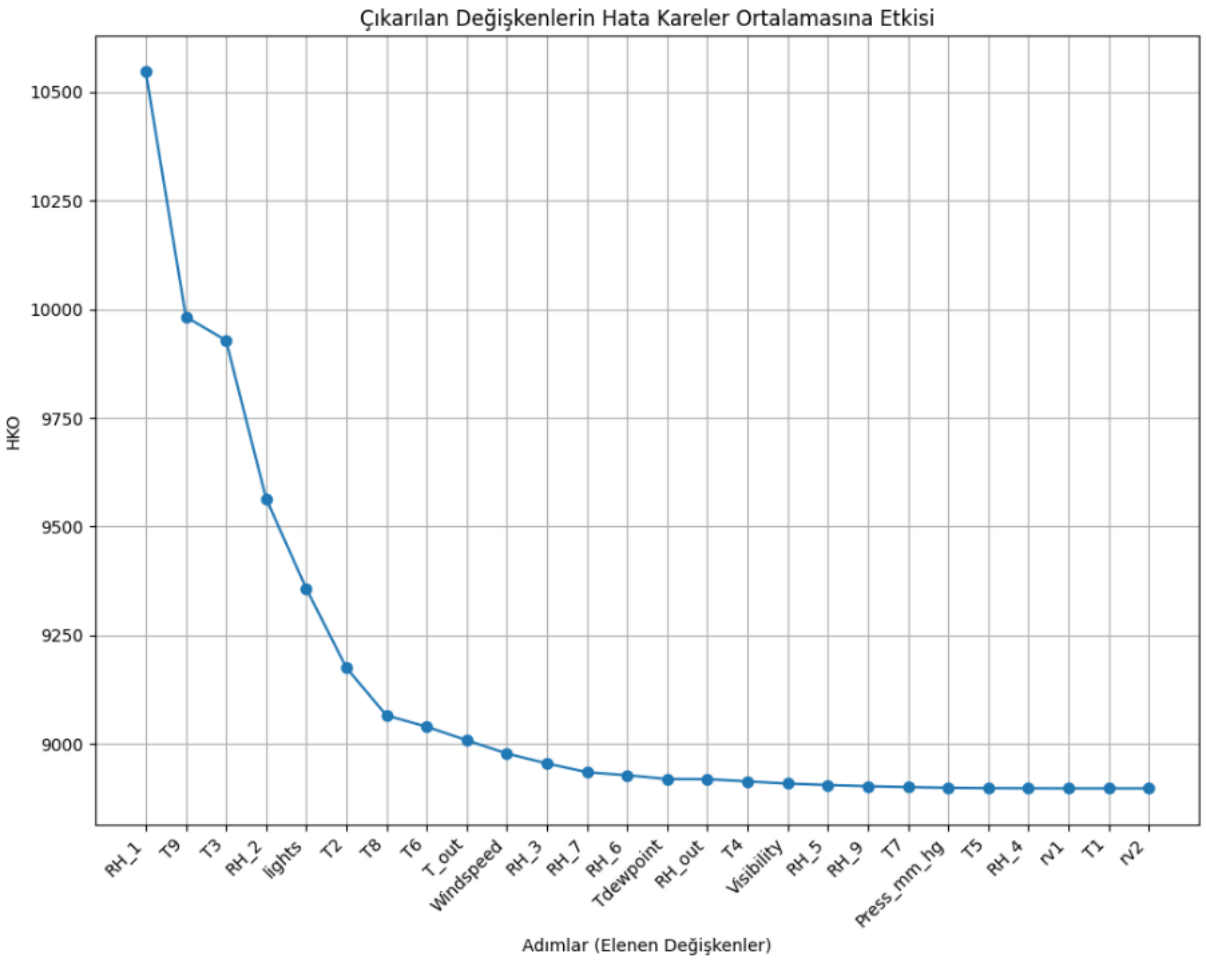
Karar Ağaçları, Destek Vektör Makineleri ve Temel Bileşenler Analizi algoritmalarıyla oluşturulan modellerin R^2 değerleri ise %10'un altında olduğu için performansları başarısızdır. Bu üç algoritmanın oluşturduğu modellerde MAE ve MSE değerleri de birbirine yakındır.

Tüm bu modellerin aksine Rastgele Ormanlar modeli diğer modellere kıyasla en iyi performansı yapan model olmuştur. Modeldeki bağımlı değişkenin, bağımsız değişkenler tarafından açıklanma oranı %51'dir ve en yüksek R^2 değerine sahiptir. MAE ve MSE değerleri de diğer modellere göre neredeyse yarı yarıya düşüktür.

İleriye Doğru Seçim yöntemi sonucunda seçilen değişkenlerle kurulan en başarılı model Rastgele Ormanlar Algoritması ile kurulan modeldir. Diğer algoritmalarla kurulan model performansları oldukça zayıf kalmıştır.

3.4.4. Geriye Doğru Eleyerek Seçim

Geriye Doğru Eleyerek Seçim yöntemi ile değişken seçimi sonucunda elenen değişkenlerin gösterilmesi amacıyla MSE değerini nasıl etkilediğinin gözlemlenebilmesi amacıyla Şekil 3.11'deki grafik elde edilmiştir.



Şekil 3.11. Değişkenlerin modele olan etkisi (MSE cinsinden)

Tüm değişkenlerin olduğu model ile başlanır ve sonra değişkenler teker teker çıkarılırken modele ait MSE değeri baz alınmıştır. Modelden çıkarılırken MSE değerini düşüren her değişken modeli olumsuz etkileyeceğinden söz konusu değişkenler modele dahil edilmemiştir. Tersten ifadesiyle, söz konusu değişkenler modelde bulunmaları halinde MSE'yi artırdığı görüldüğünden modele dahil edilmemiştir.

Şekil 3.11.'de görüldüğü üzere RH_5 değişkeninden itibaren çıkarılan değişkenlerin MSE'yi olumsuz etkisinin olmadığı görüldüğünden **RH_5, RH_9, T7, Press_mm_hg, T5, RH_4, rv_1, T1, rv2** değişkenleri seçilmiştir.

Geriye Doğru Eleyerek Seçim sonucunda ML algoritmaları ile kurulan modeller ve performans değerleri Çizelge 3.6.'te verilmiştir.

Çizelge 3.6. Geriye Doğru Eleyerek Seçim'e ilişkin sonuçlar

Algoritma/Ölçüt	**MAE**	**MSE**	**R^2**
Doğrusal Regresyon	57.81	9760.90	0.02
Karar Ağaçları	39.18	8249.81	0.18
Rastgele Ormanlar	**32.90**	**4394.68**	**0.56**
Destek Vektör Makineleri	46.59	10756.33	0.00
Temel Bileşenler Analizi	59.54	9974.80	0.01
Yapay Sinir Ağları	56.52	9338.88	0.07

Çizelge 3.6.'da ML algoritmalarıyla kurulan modeller arasında Doğrusal Regresyon modeli çok düşük bir R^2 değerine sahiptir. Bu da modeldeki bağımsız değişkenlerin bağımlı değişkeni açıklamadaki performansının çok düşük olduğunu göstermektedir. Aynı durum Temel Bileşenler Analizi ve Yapay Sinir Ağları modelleri için de geçerlidir.

Destek Vektör Makineleri model için R^2 değerleri sıfırdır. Bu algoritmalar için bağımsız değişkenlerle bağımlı değişken arasında ilişki yoktur. Bu iki model için MAE ve MSE değerleri çok yüksek çıkmıştır.

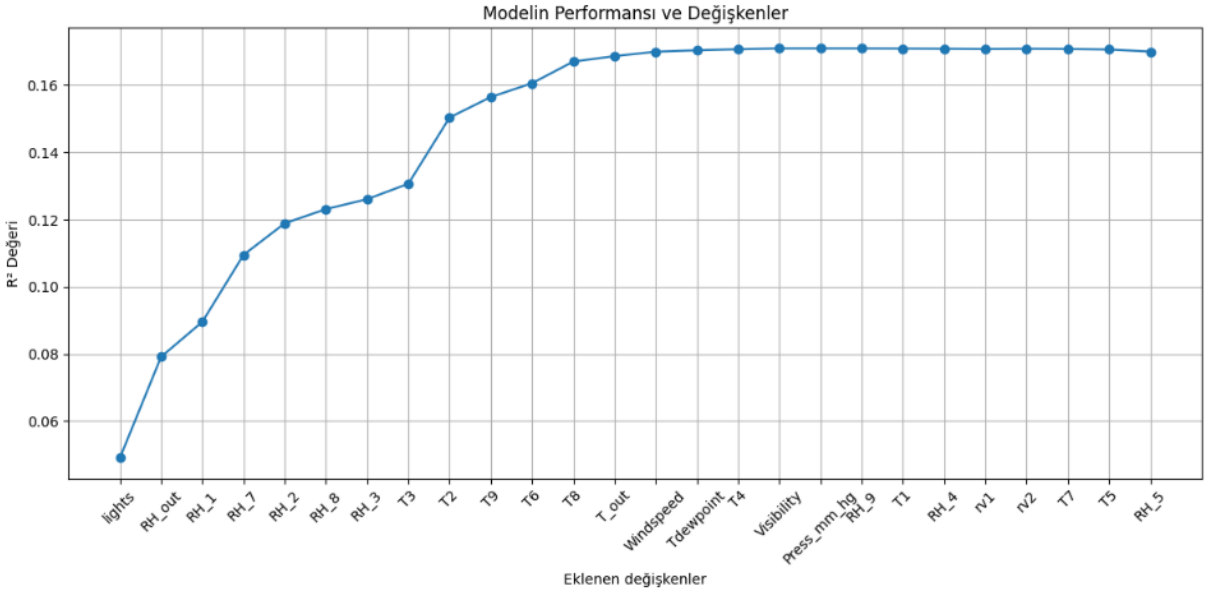
Karar Ağaçları Algoritması ile oluşturulan model ise %18'lik bir açıklama oranı elde etmiştir.

Rastgele Ormanlar modeli diğer modellere kıyasla en iyi performansı yapan model olmuştur. Modeldeki bağımlı değişkenin, sadece 9 bağımsız değişken tarafından açıklanma oranı %56'dır ve en yüksek R^2 değerine sahiptir. MAE ve MSE değerleri de diğer modellere göre neredeyse yarı yarıya düşüktür.

Geriye Doğru Eleyerek Seçim yöntemi sonucunda seçilen değişkenlerle kurulan en başarılı model Rastgele Ormanlar Algoritması ile kurulan modeldir. Diğer algoritmalarla kurulan model performansları oldukça zayıf veya geçersiz kalmıştır.

3.4.5. Adımsal Seçim

İleriye Doğru Seçim ve Geriye Doğru Eleyerek Seçim yöntemlerinin bir karışımı olan Adımsal Seçim yöntemi uygulanırken değişkenlerin MSE'ye etkisi baz alınmıştır. Bu işlem sonucunda Şekil 3.12.'de x-ekseninde seçilen tüm değişkenler gösterilmiştir ve seçilen değişkenlerin ayrıca bağımlı değişkeni ne oranda açıkladığının görülebilmesi amacıyla R^2 grafiği çizdirilmiştir.



Şekil 3.12. Değişkenlerin modele olan etkisi (R^2 cinsinden)

Şekil 3.12'ye bakıldığında **T_out** değişkeninden itibaren eklenen değişkenlerin bağımlı değişkeni açıklama oranı olan R^2'ye etkisinin olmadığı görüldüğünden **lights, RH_out, RH_1, RH_7, RH_2, RH_8, RH_3, T3, T2, T9, T6 ve T8** değişkenleri seçilmiştir.

Adımsal Seçim sonucunda ML algoritmaları ile kurulan modeller ve performans değerleri Çizelge 3.7.'te verilmiştir.

Çizelge 3.7. Adımsal Seçim'e ilişkin sonuçlar

Algoritma/Ölçüt	**MAE**	**MSE**	**R^2**
Doğrusal Regresyon	52.70	8335.55	0.17
Karar Ağaçları	41.24	9887.94	0.02
Rastgele Ormanlar	**33.21**	**4921.71**	**0.51**
Destek Vektör Makineleri	43.28	9589.31	0.04
Temel Bileşenler Analizi	54.58	8884.03	0.11
Yapay Sinir Ağları	50.03	7987.40	0.21

Çizelge 3.7'de ML algoritmalarıyla kurulan modeller arasında Karar Ağaçları ve Destek Vektör Makineleri modelleri çok düşük bir R^2 değerine sahiptir. Bu da modeldeki bağımsız değişkenlerin bağımlı değişkeni açıklamadaki performansının çok düşük olduğunu göstermektedir.

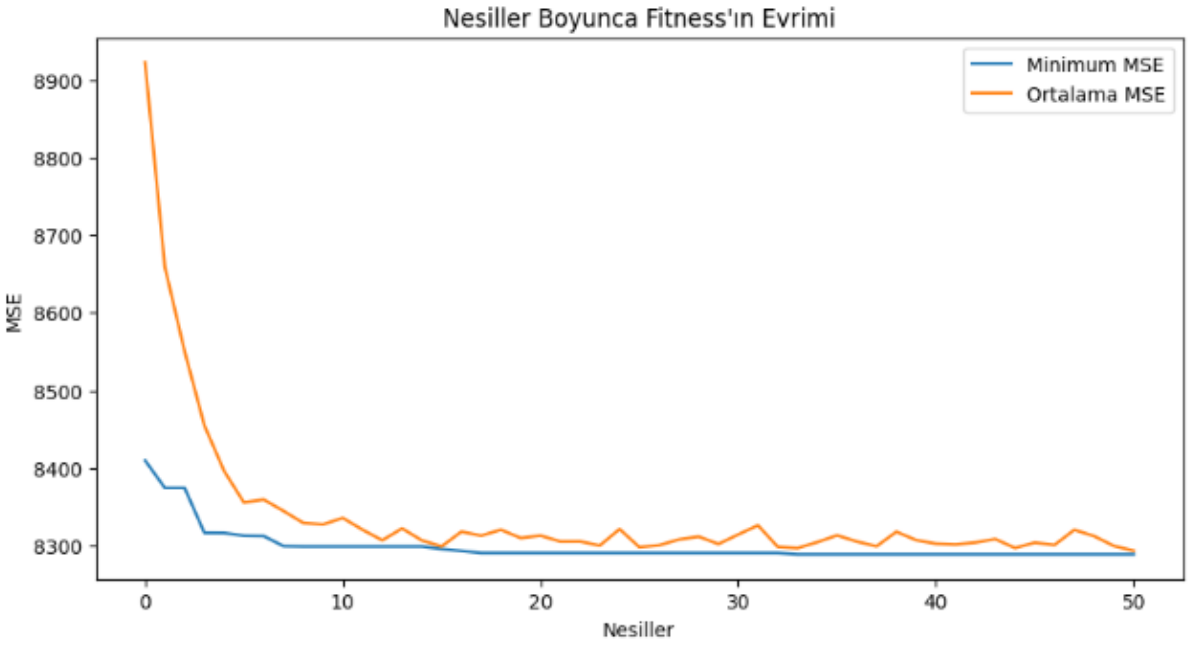
Doğrusal Regresyon, Temel Bileşenler Analizi ve Yapay Sinir Ağları algoritmaları ile oluşturulan modeller ise Karar Ağaçları ve Destek Vektör Makineleri algoritmalarına kıyasla daha iyi başarı performansı elde etmişlerdir. Yapay Sinir Ağları algoritması ile oluşturulan model %21'lik bir açıklama oranına ulaşmıştır.

Rastgele Ormanlar modeli diğer modellere kıyasla en iyi performansı elde eden model olmuştur. Modeldeki bağımlı değişkenin, bağımsız değişkenler tarafından açıklanma oranı %51'dir ve en yüksek R^2 değerine sahiptir. Çizelge3.7'de MAE ve MSE değerleri de diğer modellere göre neredeyse yarı yarıya düşüktür.

Adımsal Seçim yöntemi sonucunda seçilen değişkenlerle kurulan en başarılı model Rastgele Ormanlar Algoritması ile kurulan modeldir. Diğer algoritmalarla kurulan model performansları oldukça zayıf veya geçersiz kalmıştır.

3.4.6. Genetik Algoritmalar Tabanlı Seçim

Genetik Algoritmalar Tabanlı Seçim yöntemi ile birlikte en iyi veya en uygun değişken seçim kombinasyonları için Fitness fonksiyonu olarak MSE kullanılmıştır. Her bir kombinasyonun üretilme sürecini ifade eden Nesiller boyunca, MSE'nin nasıl etkilendiği Şekil 3.13'te gösterilmiştir.



Şekil 3.13. Oluşturulan bireylerin Nesiller boyunca MSE değerleri

Şekil 3.13'e bakıldığında model başarılı bir şekilde uygulanmış ve MSE değerini düşürmeyi başarmıştır.

Minimum MSE değerinin elde edildiği noktada değişken kombinasyonu **lights, T1, RH_1, RH_2, T3, T4, RH_4, T5, RH_7, T8, RH_8, T9, Visibility, Tdewpoint ve rv2** olarak seçilmiştir.

Genetik Algoritmalar Tabanlı Seçim sonucunda ML algoritmaları ile kurulan modeller ve performans değerleri Çizelge 3.8'te verilmiştir.

Çizelge 3.8. Genetik Algoritmalar Tabanlı Değişken Seçim'e ilişkin sonuçlar

Algoritma/Ölçüt	**MAE**	**MSE**	**R^2**
Doğrusal Regresyon	52.73	8415.47	0.16
Karar Ağaçları	42.63	9924.93	0.01
Rastgele Ormanlar	**33.92**	**4911.81**	**0.51**
Destek Vektör Makineleri	56.77	9400.36	0.06
Temel Bileşenler Analizi	43.31	9644.68	0.03
Yapay Sinir Ağları	50.66	9093.38	0.19

Çizelge 3.8.'de ML algoritmalarıyla kurulan modeller arasında Karar Ağaçları, Destek Vektör Makineleri ve Temel Bileşenler Analizi algoritmalarıyla oluşturulan modeller çok düşük bir R^2 değerine sahiptir. Bu da modeldeki bağımsız değişkenlerin bağımlı değişkeni açıklamadaki performansının çok düşük olduğunu göstermektedir.

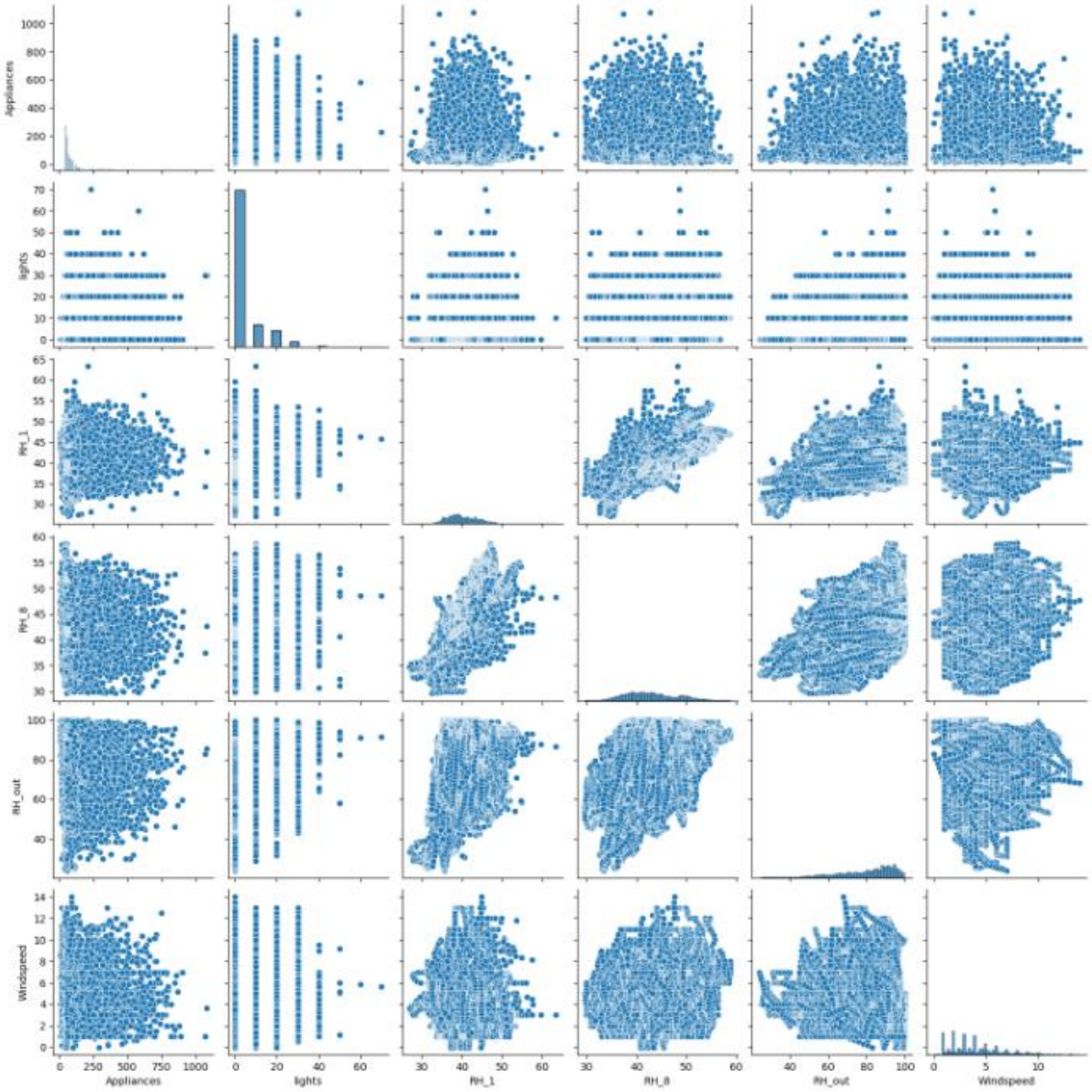
Doğrusal Regresyon ve Yapay Sinir Ağları algoritmaları ile oluşturulan modeller ise Karar Ağaçları, Destek Vektör Makineleri ve Temel Bileşenler Analizi algoritmalarına kıyasla daha iyi başarı performansı elde etmişlerdir. Doğrusal Regresyon algoritması ile oluşturulan model %16'lık açıklanma oranına ulaşırken, Yapay Sinir Ağları algoritması ile oluşturulan model %19'luk bir açıklama oranına ulaşmıştır.

Rastgele Ormanlar modeli diğer modellere kıyasla en iyi performansı elde eden model olmuştur. Modeldeki bağımlı değişkenin, bağımsız değişkenler tarafından açıklanma oranı %51'dır ve en yüksek R^2 değerine sahiptir. MAE ve MSE değerleri de diğer modellere göre neredeyse yarı yarıya düşüktür.

Genetik Algoritmalar Tabanlı Değişken Seçim yöntemi ile seçilen değişkenlerle kurulan en başarılı model Rastgele Ormanlar Algoritması modelidir. Diğer algoritmalarla kurulan model performansları oldukça zayıf veya geçersiz kalmıştır.

3.4.7. Lasso Regresyon Tabanlı Seçim

Değişken katsayılarını küçülterek en iyi veri kümesini bulmaya çalışan Lasso Regresyon Tabanlı Yöntem ile yapılan değişken seçimi sonucunda **lights, RH_1, RH_8, RH_out, Windspeed** değişkenleri seçilmiştir. Seçilen değişkenlerin birbirleriyle olan ilişkisini gösteren saçılım grafiği Şekil 3.14.'te verilmiştir.



Şekil 3.14. Lasso Regresyon ile seçilen değişkenler için saçılım grafiği

Şekil 3.14'e bakıldığında bağımlı değişken ile bağımsız değişkenler arasındaki ilişkinin rastgele olduğu görülmektedir. RH_8 ile RH_1 değişkeni arasında pozitif ve doğrusal bir ilişki vardır.

Lasso Regresyon Tabanlı Seçim sonucunda ML algoritmaları ile kurulan modeller ve performans değerleri Çizelge 3.9.'te verilmiştir.

Çizelge 3.9. Lasso Regresyon Tabanlı Seçim'e ilişkin sonuçlar

Algoritma/Ölçüt	**MAE**	**MSE**	**R^2**
Doğrusal Regresyon	54.58	8890.12	0.11
Karar Ağaçları	50.04	12202.24	0.00
Rastgele Ormanlar	**42.38**	**6714.69**	**0.33**
Destek Vektör Makineleri	45.24	9956.69	0.01
Temel Bileşenler Analizi	54.58	8890.12	0.11
Yapay Sinir Ağları	53.76	8978.58	0.11

Çizelge 3.9.'da ML algoritmalarıyla kurulan modeller arasında Destek Vektör Makineleri algoritması ile oluşturulan model çok düşük bir R^2 değerine sahiptir. Bu da modeldeki bağımsız değişkenlerin bağımlı değişkeni açıklamadaki performansının çok düşük olduğunu göstermektedir.

Doğrusal Regresyon, Temel Bileşenler Analizi ve Yapay Sinir Ağları algoritmaları ile oluşturulan modeller ise Destek Vektör Makineleri algoritmasına kıyasla daha iyi başarı performansı elde etmişlerdir. Doğrusal Regresyon, Temel Bileşenler Analizi ve Yapay Sinir Ağları algoritmaları ile oluşturulan modeller %11'lik bir açıklama oranına sahiptir.

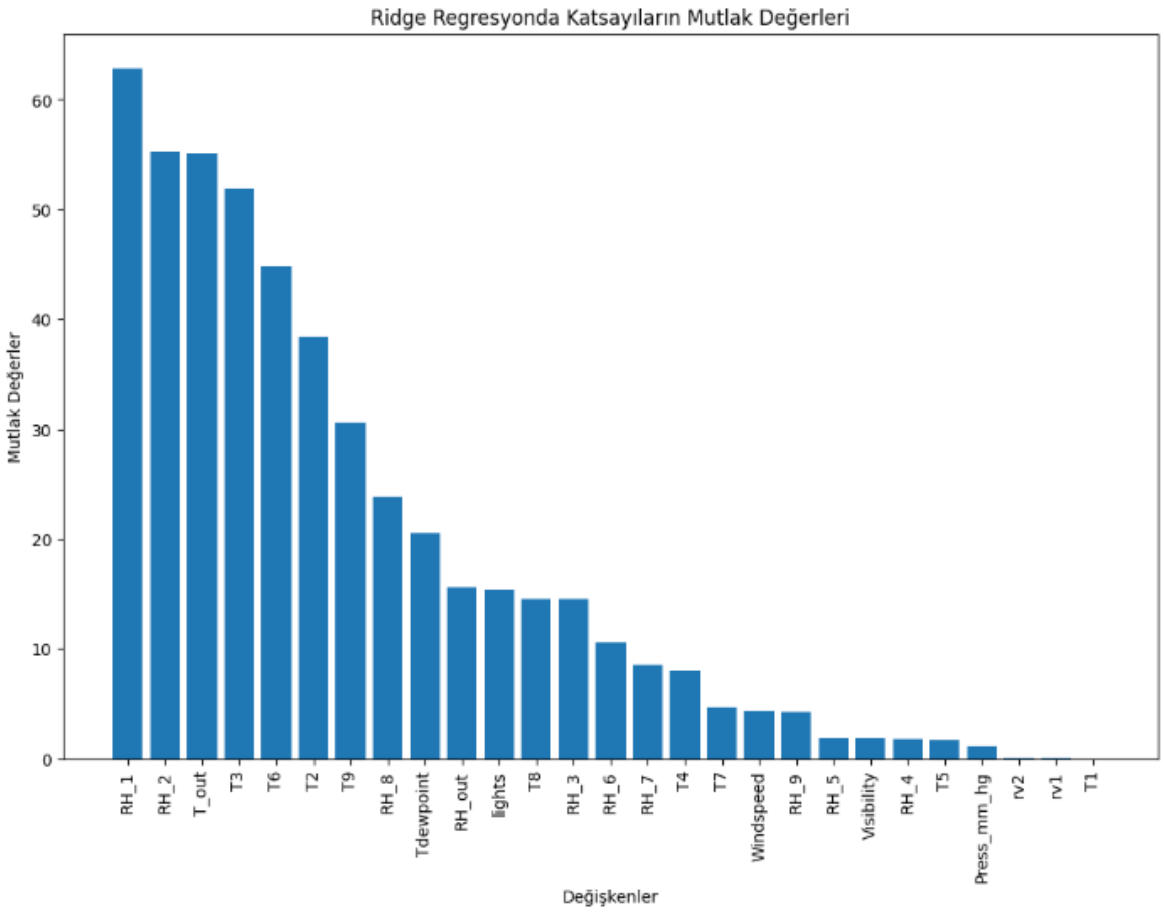
Karar Ağaçları modeli için R^2 değerleri sıfırdır. Bu algoritmalar için bağımsız değişkenlerle bağımlı değişken arasında ilişki yoktur. Bu iki model için MAE ve MSE değerleri çok yüksek çıkmıştır.

Rastgele Ormanlar modeli diğer modellere kıyasla en iyi performansa sahip model olmuştur. Modeldeki bağımlı değişkenin, bağımsız değişkenler tarafından açıklanma oranı %33'tür ve en yüksek R^2 değerine sahiptir. MAE ve MSE değerleri de diğer modellere göre neredeyse yarı yarıya düşüktür. Bu anlamda da model seçim kriterini sağlamaktadır. Diğer algoritmalara kıyasla üstün performanslı olmasına rağmen başarı değeri düşüktür.

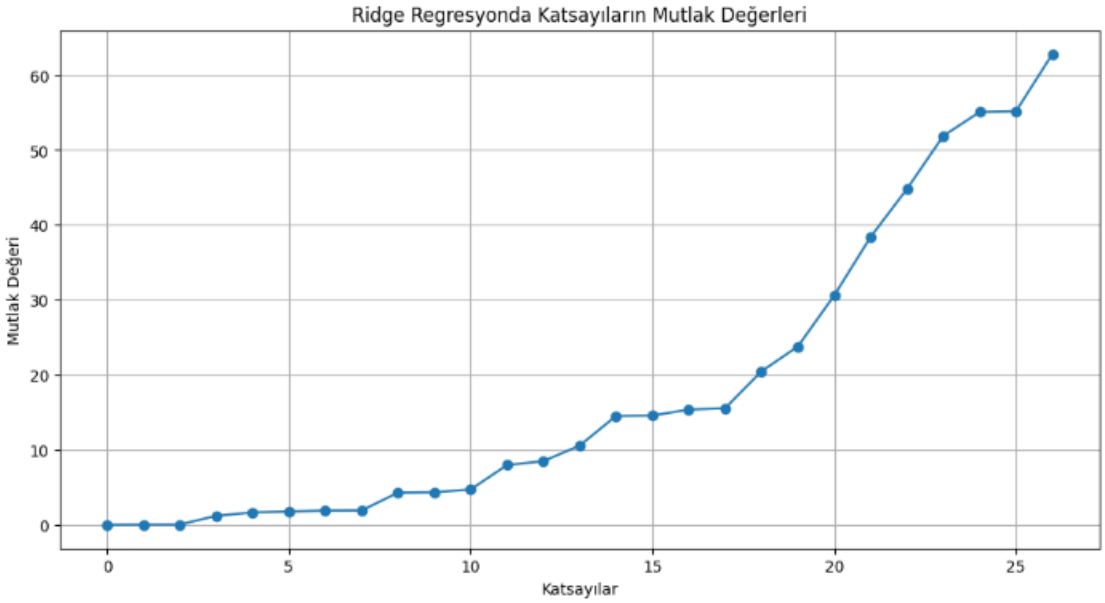
Lasso Regresyon Tabanlı Değişken Seçim yöntemi sonucunda seçilen değişkenlerle kurulan performansı en yüksek model Rastgele Ormanlar Algoritması ile kurulan modeldir. Diğer algoritmalarla kurulan model performansları oldukça zayıf veya geçersiz kalmıştır.

3.4.8. Ridge Regresyon Tabanlı Seçim

Ridge Regresyon Tabanlı Seçim yöntemi ile değişken seçimi yapılırken katsayıları en yüksek olan değişkenler seçilmiştir. Bunun yanında kırılma noktaları, yani eğimleri de göz önünde bulundurulmuştur. Kırılımı az ve katsayısı düşük olan değişkenler modelden çıkarılmıştır. Şekil 3.15. ve Şekil 3.16.'te değişkenlerin katsayılarının mutlak değerleriyle oluşturulmuş grafikler verilmiştir.



Şekil 3.15. Değişkenlerin katsayılarının mutlak değerleri için çubuk grafiği



Şekil 3.16. Katsayıların mutlak değeri için çizgi grafiği

Ridge Regresyon Tabanlı Seçim yöntemi uygulanması sonucunda **RH_1, RH_2, T3, T6, T2, T9, RH_8 ve Tdewpoint** değişkenleri seçilmiştir. Zira bu değişkenler katsayıları en yüksek değişkenlerdir. Bunun yanında Şekil 3.16.'daki eğimi en çok bu değişkenler yaratmaktadır.

Ridge Regresyon Tabanlı Seçim sonucunda seçilen değişkenlerle oluşturulan ML algoritmaları ile kurulan modeller ve performans değerleri Çizelge 3.10.'da verilmiştir.

Çizelge 3.10. Ridge Regresyon Tabanlı Seçim'e ilişkin sonuçlar

Algoritma/Ölçüt	**MAE**	**MSE**	**R^2**
Doğrusal Regresyon	54.1	8746.24	0.12
Karar Ağaçları	38.43	8608.18	0.14
Rastgele Ormanlar	**31.78**	**4450.89**	**0.55**
Destek Vektör Makineleri	45.34	9879.46	0.01
Temel Bileşenler Analizi	55.54	9278.51	0.07
Yapay Sinir Ağları	52.45	8369.45	0.16

Çizelge 3.10'da ML algoritmalarıyla kurulan modeller arasında Destek Vektör Makineleri ve Temel Bileşenler Analizi algoritmaları ile oluşturulan modeller çok düşük bir R^2 değerine sahiptir. Bu da modeldeki bağımsız değişkenlerin bağımlı değişkeni açıklamadaki performansının çok düşük olduğunu göstermektedir.

Doğrusal Regresyon, Karar Ağaçları ve Yapay Sinir Ağları algoritmaları ile oluşturulan modeller ise Destek Vektör Makineleri algoritmasına kıyasla daha iyi başarı performansı elde etmişlerdir. Doğrusal Regresyon, Temel Bileşenler Analizi ve Yapay Sinir Ağları algoritmaları ile oluşturulan modeller sırasıyla %12, %14 ve %16'lık bir açıklanma oranına sahiptir.

Rastgele Ormanlar modeli diğer modellere kıyasla en iyi başarı performansı elde eden model olmuştur. Modeldeki bağımlı değişkenin, sadece 8 bağımsız değişken tarafından açıklanma oranı %55'tir ve en yüksek R^2 değerine sahiptir. MAE ve MSE değerleri de diğer modellere göre neredeyse yarı yarıya düşüktür.

Ridge Regresyon Tabanlı Değişken Seçim yöntemi sonucunda seçilen değişkenlerle kurulan modeller arasında performansı en yüksek olan model Rastgele Ormanlar Algoritması ile kurulan modeldir. Diğer algoritmalarla kurulan model performansları oldukça zayıf kalmıştır.

3.4.9. Sağlam Değişken Seçim Yöntemi

Şekil 3.2'den verinin normal dağılmadığı belirtilmişti. Veriye ilişkin öninceleme yapıldığında aykırı değerlerin olduğu görülmüştür. Ayrıca çoklu bağlantı sorununun olup olmadığı incelenmiş ve değişkenlere ilişkin VIF (Varyans Şişme Katsayıları) değerleri Çizelge 3.11'de verilmiştir. T6, T_out, RH_out ve Tdewpoint değişkenlerine ilişkin VIF değerleri 30'dan büyük olduğundan çoklu bağlantı sorunu vardır diyebiliyoruz. Aykırı değer ve çoklu bağlantı sorunu birlikte görüldüğünden Sağlam Değişken Seçim Yöntemi kullanılmıştır.

Çizelge 3.11. Değişkenlere ilişkin VIF değerleri

Değişken	VIF Değeri
Visibility	1.04
lights	1.28
RH_5	1.37
Press_mm_hg	1.41
Windspeed	1.61
RH_9	6.48
T8	8.05
RH_8	8.51
T4	9.83
RH_6	9.99
T3	10.00
T5	10.51
RH_3	10.82

Değişken	VIF Değeri
RH_7	10.82
RH_1	15.96
RH_4	17.12
T7	17.50
T1	19.67
RH_2	21.89
T9	28.27
T2	28.80
T6	**33.46**
RH_out	**49.23**
Tdewpoint	**86.12**
T_out	**146.84**

ML algoritmalarının en çok seçtiği 15 değişken ve bu değişkenlerin kullanılma sıklıkları Çizelge 3.12'de verilmiştir.

Çizelge 3.12. En çok kullanılan 15 değişkenin kullanım sıklıklar

Değişken	**Frekans**
lights	5
RH_1	5
RH_8	5
T2	4
T9	4
T3	4
RH_2	4
T8	3
RH_7	3
Press_mm_hg	2
Windspeed	2
T5	2
T6	5
RH_out	4
Tdewpoint	3

VIF değeri 30'dan yüksek olan değişkenler T6, RH_out ve Tdewpoint değişkenleri veriden çıkartılarak 12 bağımsız değişkene ilişkin sağlam seçim ölçütleri hesaplanmıştır. AIC, AICC, BIC ve BICC seçim ölçütlerinde klasik kestirim değerleri yerine Sağlam Huber M-kestirim değerleri kullanılmış ve 2^{12}-1= 4095 altküme için bu seçim yöntemlerinin sağlam versiyonları hesaplanmıştır. 4095 tane alt küme içinde en düşük AIC ve BIC değerlerine sahip değişken kümesi **lights, RH_1, RH_8, T2, T3, RH_2, RH_7, Press_mm_hg, Windspeed ve T5** değişkenlerini içermektedir. T9 veT8 değişkenleri modele girememiştir.

Sağlam Değişken Seçim Yöntemi tarafından seçilen bağımsız değişkenlerle kurulan ML algoritmaları ve bu algoritmalara ilişkin performans sonuçları Çizelge 3.13'te verilmiştir.

Çizelge 3.13. Sağlam Değişken Seçim Yöntem'ine ilişkin sonuçlar

Algoritma/Ölçüt	MAE	MSE	R^2
Doğrusal Regresyon	53.81	8528.22	0.15
Karar Ağaçları	39.53	8882.92	0.11
Rastgele Ormanlar	**33.45**	**4901.15**	**0.51**
Destek Vektör Makineleri	44.43	9732.89	0.03
Temel Bileşenler Analizi	55.97	9261.41	0.07
Yapay Sinir Ağları	51.57	8256.16	0.17

Çizelge 3.13'te ML algoritmalarıyla kurulan modeller arasında Destek Vektör Makineleri ve Temel Bileşenler Analizi algoritmaları ile oluşturulan modeller çok düşük bir R^2 değerine sahiptir. Bu da modeldeki bağımsız değişkenlerin bağımlı değişkeni açıklamadaki performansının çok düşük olduğunu göstermektedir.

Doğrusal Regresyon, Karar Ağaçları ve Yapay Sinir Ağları algoritmaları ile oluşturulan modeller ise Destek Vektör Makineleri algoritmasına kıyasla daha iyi başarı performansı elde etmişlerdir. Doğrusal Regresyon, Temel Bileşenler Analizi ve Yapay Sinir Ağları algoritmaları ile oluşturulan modeller sırasıyla %15, %11 ve %17'lik bir açıklama oranına sahiptir.

Rastgele Ormanlar modeli diğer modellere kıyasla en iyi başarı performansı elde eden model olmuştur. Açıklama oranı %51'dir ve en yüksek R^2 değerine sahiptir. MAE ve MSE değerleri de diğer modellere göre neredeyse yarı yarıya düşüktür.

Sağlam Değişken Seçim Yöntemlerinin uygulanması sonucunda seçilen değişkenlerle kurulan modeller arasında performansı en yüksek olan model Rastgele Ormanlar Algoritması ile kurulan modeldir. Diğer algoritmalarla kurulan model performansları oldukça zayıf kalmıştır.

4. SONUÇ

Bu çalışmada, enerji tüketimi tahmininde kullanılan değişken seçim yöntemlerinin farklı ML algoritmaları üzerindeki etkisi detaylı bir şekilde incelenmiştir. Çalışmanın temel amacı, değişken seçim yöntemlerinin ML algoritmalarının performansına olan etkilerini değerlendirmek ve hangi kombinasyonların en iyi sonuçları verdiğini belirlemektir. Bu kapsamda, değişken seçim yöntemleri olarak Korelasyon Tabanlı Seçim, Varyans Tabanlı Seçim, İleriye Doğru Seçim, Geriye Doğru Eleyerek Seçim, Adımsal Seçim, Genetik Algoritmalar Tabanlı Seçim, Lasso Regresyonu Tabanlı Seçim, Ridge Regresyon Tabanlı Seçim ve Sağlam Değişken Seçim Yöntemleri kullanılmış ve ML algoritmaları olarak Doğrusal Regresyon, Karar Ağaçları, Rastgele Ormanlar, Destek Vektör Makineleri, Temel Bileşenler Analizi ve Yapay Sinir Ağları kullanılarak modeller oluşturulmuş ve performansları karşılaştırılmıştır.

Çalışmanın bulguları, değişken seçim yöntemleri ve ML algoritmalarının enerji tüketimi performansı üzerinde önemli etkiler yarattığını ortaya koymuştur. Özellikle Rastgele Ormanlar algoritması hem Geriye Doğru Eleyerek Seçim hem de Ridge Regresyon Tabanlı Seçim yöntemleriyle en iyi performansı sergilemiştir. Rastgele Ormanlar algoritması, en düşük MAE ve MSE değerleri ile en yüksek R^2 değerine ulaşarak verilerin önemli bir kısmını açıklamış ve tahmin doğruluğunu artırmıştır.

Bu durum, Rastgele Ormanlar algoritmasının esnekliği ve yüksek tahmin doğruluğu sağlamadaki yeteneğini göstermektedir. Diğer yandan, Karar Ağaçları ve Temel Bileşenler Analizi algoritmaları, düşük performans sergileyerek enerji tüketimi tahmininde yetersiz kalmıştır. Değişken Seçim Yöntemi olarak ise Lasso Regresyon Tabanlı Seçim yöntemi en düşük performansı sergilemiştir.

Veri kümesi, normal dağılıma uygun bir dağılım sergilememekte ve aykırı değerler içermektedir. Ayrıca, VIF değeri 30'un üzerinde olan değişkenlerin neden olduğu çoklu bağlantı sorunu bulunduğundan, Sağlam Değişken Seçim Yöntemi uygulanmıştır. Uygulanan 9 farklı

değişken seçim yöntemlerince (Korelasyon Tabanlı Seçim, Varyans Tabanlı Seçim, İleriye Doğru Seçim, Geriye Doğru Eleyerek Seçim, Adımsal Seçim, Genetik Algoritmalar Tabanlı Seçim, Lasso Regresyonu Tabanlı Seçim ve Ridge Regresyon Tabanlı Seçim) en çok seçilen 15 değişken belirlenmiştir. VIF değerleri 30'dan yüksek olan değişkenler çıkarıldıktan sonra geriye kalan 12 değişkene Sağlam Değişken Seçim Yöntemi uygulanmıştır. Toplamda 4095 değişken kombinasyonu arasından en düşük AIC, AICC, BIC ve BICC değerine sahip değişkenler seçilmiş ve ML algoritmaları uygulanması sonucu performansları ölçülmüştür. Sağlam Değişken Seçim Yöntemi uygulanması sonucunda en başarılı performansa sahip olan ML algoritması Rastgele Ormanlar modeli olmuştur.

Seçilen değişkenlerin kullanım sıklıkları incelendiğinde en çok kullanılan değişkenler lights (evde kullanılan aydınlatmanın enerji kullanım miktarı, RH_1 (mutfak alanındaki nem) ve RH_8 (genç odasındaki nem) değişkenleridir. Bu değişkenler uygulanan 9 değişken seçim yönteminin 6'sında değişken seçim yöntemleri tarafından seçilen değişken kümelerinde mevcuttur. Bunun yanında dış sıcaklığı temsil eden T_out (dış sıcaklık) ve binanın dış cephesindeki nemi temsil eden RH_6 (binanın dışındaki nem) değişkenleri hiçbir değişken seçim yöntemi uygulaması sonucunda seçilmemiş ve modellere dahil edilmemiştir.

Bu çalışma, enerji tüketimi tahmininde doğru değişken seçimi ve uygun ML algoritmalarının kullanılmasının önemini vurgulamaktadır. Elde edilen sonuçlar, enerji tüketimi tahmin modellerinin doğruluğunu artırmak için hangi değişken seçim yöntemlerinin ve ML algoritmalarının daha etkili olduğunu göstermektedir. Bu bulgular, veri bilimcileri ve araştırmacılar için değerli bilgiler sunarak veri kümesine uygun yöntem ve algoritma seçiminde rehberlik etmektedir.

Gelecekteki araştırmalar, farklı veri kümeleri ve coğrafi bölgeler üzerinde benzer analizlerin yapılmasını içerdiği takdirde bu sonuçların genelleştirilebilirliğini artıracaktır. Ayrıca, daha geniş bir ML algoritması ve hiper-parametre optimizasyon teknikleri yelpazesi kullanılarak modellerin performansının artırılması mümkündür. Diğer tahmin problemleri üzerinde değişken seçim yöntemlerinin etkisinin incelenmesi de önemli bir araştırma alanı olabilir.

Özellikle, farklı endüstrilerdeki uygulamaların incelenmesi, değişken seçim yöntemlerinin ve ML algoritmalarının genel geçer etkinliğini değerlendirmeye yardımcı olabilir.

Sonuç olarak, bu çalışma, enerji tüketimi tahmininde Değişken Seçim Yöntemleri'nin ve ML algoritmalarının etkinliğini karşılaştırmalı olarak ortaya koymuştur. Bulgular, enerji tüketimi tahmin modellerinin doğruluğunu artırmak için uygun değişken seçim yöntemlerinin ve ML algoritmalarının önemini vurgulamaktadır. Bu alanda yapılan araştırmaların, enerji verimliliği stratejilerinin geliştirilmesine ve sürdürülebilir enerji yönetimi politikalarına katkı sağlaması beklenmektedir.

5. KAYNAKLAR

[1] M. Buyukkececi, M.C. Okur, A Comprehensive Review of Feature Selection and Feature Selection Stability in Machine Learning, Gazi University Journal of Science 36 (**2023**) 1506–1520.

[2] G. Chandrashekar, F. Sahin, A survey on feature selection methods, Computers and Electrical Engineering 40 (**2014**) 16–28.

[3] N. Pudjihartono, T. Fadason, A.W. Kempa-Liehr, J.M. O'Sullivan, A Review of Feature Selection Methods for Machine Learning-Based Disease Risk Prediction, Frontiers in Bioinformatics 2 (**2022**).

[4] J.K. Sethi, M. Mittal, A new feature selection method based on machine learning technique for air quality dataset, Journal of Statistics and Management Systems 22 (**2019**) 697–705.

[5] I.D. Mienye, Y. Sun, A Machine Learning Method with Hybrid Feature Selection for Improved Credit Card Fraud Detection, Applied Sciences (Switzerland) 13 (**2023**).

[6] C. Gazeloğlu, Prediction of heart disease by classifying with feature selection and machine learning methods, Progress in Nutrition 22 (**2020**) 660–670.

[7] N. Challita, M. Khalil, P. Beauseroy, New feature Selection method based on neural network and machine learning, in: IEEE International Multidisciplinary Conference on Engineering Technology (IMCET), **2016**.

[8] F. Saeed, M. Al-Sarem, M. Al-Mohaimeed, A. Emara, W. Boulila, M. Alasli, F. Ghabban, Enhancing Parkinson's Disease Prediction Using Machine Learning and Feature Selection Methods, Computers, Materials and Continua 71 (**2022**) 5639–5657.

[9] R.C. Chen, C. Dewi, S.W. Huang, R.E. Caraka, Selecting critical features for data classification based on machine learning methods, J Big Data 7 (**2020**).

[10] M. Ünlüsavuran, Omik Verilerinde Otomatik Makine Öğrenimi Algoritmalarının Performansının Değerlerndirilmesi, Yüksek Lisans Tezi, Erciyes Üniversitesi Sağlık Bilimleri Enstitüsü, **2019**.

[11] M. Çetin, Sağlam Regresyonda Değişken Seçim Ölçütleri, Doktora Tezi, Hacettepe Üniversitesi Fen Bilimleri Enstitüsü, **2000**.

[12] O. Toka, M. Çetin, O. Arslan, Robust regression estimation and variable selection when cellwise and casewise outliers are present, Hacettepe Journal of Mathematics and Statistics 50 (**2021**) 289–303.

[13] N. Pudjihartono, T. Fadason, A.W. Kempa-Liehr, J.M. O'Sullivan, A Review of Feature Selection Methods for Machine Learning-Based Disease Risk Prediction, Frontiers in Bioinformatics 2 (**2022**).

[14] N. Bulut, S. Shakeri, S. Yüzük, M. Aktaş, Öznitelik Yöntemleri ve Makine Öğrenmesi Kullanarak Şirket Bilanço Verilerine Dayalı İflas Riski Tahmini, Türkiye Bilişim Vakfı Bilgisayar Bilimleri ve Mühendisliği Dergisi 12 (**2019**) 20–29.

[15] O. Kaynar, H. Arslan, Y. Görmez, Y.E. Işık, Makine Öğrenmesi ve Öznitelik Seçim Yöntemleriyle Saldırı Tespiti, Bilişim Teknolojileri Dergisi 11 (**2018**) 175–185.

[16] T. Kaynar, Ö.E. Yiğit, Öznitelik Mühendisliği ile Makine Öğrenmesi Yöntemleri Kullanılarak BIST 100 Endeksi Değişiminin Tahminine Yönelik Bir Yaklaşım, Journal of Yasar University (**2021**) 1741–1762.

[17] A. Turing, Computing Machinery and Intelligence, Source: Mind, New Series 59 (**1950**) 433–460.

[18] A. Samuel, Some Studies in Machine Learning Using the Game of Checkers, IBM Journal (**1959**) 211–229.

[19] C. Arf, Makine Düşünebilir Mi ve Nasıl Düşünebilir?, in: Üniversite Çalışmalarını Muhite Yayma ve Halk Eğitimi Yayınları Konferans Serisi No:1, Erzurum, **1959**: pp. 91–103.

[20] G. Dejong, Explanation-Based Learning: An Alternative View, Mach Learn 1 (**1986**) 145–176.

[21] J.H. Holland, Genetic Algorithms, Sci Am 267 (**1992**) 66–73.

[22] D.E. Rumelhart, G.E. Hinton, R.J. Williams, Learning Representations by Back-propagating Errors, Nature 323 (**1986**) 533–536.

[23] C. Cortes, V. Vapnik, L. Saitta, Support-Vector Networks Editor, Machine Leaming 20 (**1995**) 273–297.

[24] A. Krizhevsky, I. Sutskever, G.E. Hinton, ImageNet Classification with Deep Convolutional Neural Networks, Commun ACM (**2017**) 84–90.

[25] I. Goodfellow, Y. Bengio, A. Courville, Deep Learning, MIT, **2017**.

[26] P. Domingos, A Few Useful Things to Know About Machine Learning, Commun ACM 55 (**2012**) 78–87.

[27] A. Halevy, P. Norving, F. Pereira, The Unreasonable Effectiveness of Data, IEEE (**2009**) 8–12.

[28] S. García, J. Luengo, F. Herrera, Data Preprocessing in Data Mining, Springer, Switzerland, **2015**.

[29] A. Zheng, A. Casari, Feature Engineering for Machine Learning, 1st ed., O'Reilly, Sebastopol, **2018**.

[30] T. Hastie, R. Tibshirani, J. Friedman, The Elements of Statistical Learning: Data Mining, Inference, and Prediction, 2nd ed., **2009**.

[31] N.R. Draper, H. Smith, Applied Regression Analysis, 3rd ed., John Wiley&Sons, New York, **1998**.

[32] D.W. Hosmer, Stanley. Lemeshow, R.X. Sturdivant, Applied Logistic Regression, 3rd ed., John Wiley & Sons, New Jersey, **2013**.

[33] L. Breiman, J. Friedman, R. Olshen, C. Stone, Classification and Regression Trees, Chapman & Hall, Boca Raton, **1984**.

[34] L. Breiman, Random Forests, Mach Learn 45 (**2001**) 5–32.

[35] A.J. Smola, B. Sch¨olkopf, S. Sch¨olkopf, A tutorial on support vector regression, Stat Comput 14 (**2004**) 199–222.

[36] W. Samek, T. Wiegand, K.-R. Müller, Explainable Artificial Intelligence: Understanding, Visualizing and Interpreting Deep Learning Models, (**2017**).

[37] J. Macqueen, Some Methods For Classification and Analysis of Multivariate Observations, in: Fifth Berkeley Symposium, **1967**: pp. 281–297.

[38] İ. Arslan, Python ile Veri Bilimi, 4th ed., Pusula, İstanbul, **2021**.

[39] L. Rokach, O. Maimon, Clustering Methods, Data Mining and Knowledge Discovery Handbook (**2006**) 321–352.

[40] I. Jolliffe, Principal Component Analysis Second Edition, 2nd ed., Springer, New York, **2002**.

[41] I.T. Jolliffe, A Note on the Use of Principal Components in Regression, Source: Journal of the Royal Statistical Society. Series C (Applied Statistics) 31 (**1982**) 300–303.

[42] D. Yarowsky, Unsupervised Word Sense Disambiguation Rivaling Supervised Methods, in: 33rd Annual Meeting of the Association for Computational Linguistics, **1995**: pp. 189–196.

[43] A. Blum, T. Mitchell, Combining Labeled and Unlabeled Data with Co-Training*, in: Proceedings of the 11th Annual Conference on Computational Learning Theory, **1998**: pp. 92–100.

[44] R.S. Sutton, A.G. Barto, Reinforcement Learning: An Introduction, MIT Press, London, **1998**.

[45] G.H. John, R. Kohavi, K. Pfleger, Irrelevant Features and the Subset Selection Problem, Morgan Kaufmann Publishers, **1994**.

[46] I. Guyon, A. Elisseeff, An Introduction to Variable and Feature Selection André Elisseeff, Journal of Machine Learning Research 3 (**2003**) 1157–1182.

[47] D. Koller, M. Sahami, Toward Optimal Feature Selection, in: Proceedings of the Thirteenth International Conference on International Conference on Machine Learning, **1996**: pp. 284–292.

[48] J. Brownlee, Data Preparation for Machine Learning, 1st ed., Machine Learning Mastery, **2020**.

[49] Y. Saeys, I. Inza, P. Larrañaga, A Review of Feature Selection Techniques in Bioinformatics, Bioinformatics 23 (**2007**) 2507–2517.

[50] J.H. (John H. Holland, Adaptation in Natural and Artificial Systems, MIT Press, **1992**.

[51] R. Tibshirani, Regression Shrinkage and Selection via the Lasso, Journal of the Royal Statistical Society. Series B (Methodological) 58 (**1996**) 267–288.

[52] A.E. Hoerl, R.W. Kennard, Ridge Regression: Biased Estimation for Nonorthogonal Problems, Technometrics 12 (**1970**) 55–67.

[53] E. Ronchetti, Robust Model Selection in Regression, Stat Probab Lett 3 (**1985**) 21–23.

[54] M. Çetin, A. Erar, Sağlam Cp Kriterinin Uygulanması Üzerine Bir Çalışma, Anadolu Üniversitesi Bilim ve Teknoloji Dergisi 1 (**2000**) 77–84.

[55] H. Akaike, A New Look at the Statistical Model Identification, IEEE Trans Automat Contr (**1974**).

[56] R.E. Kass, A.E. Raftery, Bayes factors, J Am Stat Assoc 90 (**1995**) 773–795.

[57] C.M. Hurvich, C.-L. Tsai, Regression and time series model selection in small samples, Biometrika 76 (**1989**) 297–307.

[58] L.M. Candanedo, V. Feldheim, D. Deramaix, Data driven prediction models of energy use of appliances in a low-energy house, Energy Build 140 (**2017**) 81–97.

[59] Appliances Energy Prediction. https://archive.ics.uci.edu/dataset/374/appliances+energy+prediction (erişim tarihi **29 Şubat, 2024**).

[60] Spring 2016 Weather History at Chièvres Air Base. https://weatherspark.com/h/s/147966/2016/0/Historical-Weather-Spring-2016-at-Chi%C3%A8vres-Air-Base-Belgium#Figures-Temperature (erişim tarihi **14 Mayıs, 2024**).