

**T.C.
MUĞLA SITKI KOÇMAN ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ**

ÇEVRE BİLİMLERİ ANABİLİM DALI

**METEOROLOJİK ÖLÇÜM VERİLERİNİN
REGRESYON YAKLAŞIMI İLE MODELLENMESİNDE
VERİ MADENCİLİĞİ AŞAMALARI VE KULLANIMI**

YÜKSEK LİSANS TEZİ

SELÇUK KURT

OCAK 2024

MUĞLA

T.C.
MUĞLA SITKI KOÇMAN ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

ÇEVRE BİLİMLERİ ANABİLİM DALI

METEOROLOJİK ÖLÇÜM VERİLERİNİN
REGRESYON YAKLAŞIMI İLE MODELLENMESİNDE
VERİ MADENCİLİĞİ AŞAMALARI VE KULLANIMI

YÜKSEK LİSANS TEZİ

SELÇUK KURT

OCAK 2024

MUĞLA

MUGLA SITKI KOÇMAN ÜNİVERSİTESİ
Fen Bilimleri Enstitüsü

TEZ ONAYI

SELÇUK KURT tarafından hazırlanan **METEOROLOJİK ÖLÇÜM VERİLERİNİN REGRESYON YAKLAŞIMI İLE MODELLENMESİNDE VERİ MADENCİLİĞİ AŞAMALARI VE KULLANIMI** başlıklı tezinin, 16/01/2024 tarihinde aşağıdaki jüri tarafından Çevre Bilimleri Anabilim Dalı'nda yüksek lisans derecesi için gerekli şartları sağladığı oybirliği/oyçokluğu ile kabul edilmiştir.

TEZ SINAV JURİSİ

Prof. Dr. Fatma AYHAN (Jüri Başkanı)
Kimya Anabilim Dalı,
Muğla Sıtkı Koçman Üniversitesi, Muğla

İmza:

Doç. Dr. Oğuz AKPOLAT (Danışman)
Çevre Bilimleri Anabilim Dalı,
Muğla Sıtkı Koçman Üniversitesi, Muğla

İmza:

Dr. Ceyhun KAPUCU (İkinci danışman)
Enformatik Bölüm Başkanlığı,
Muğla Sıtkı Koçman Üniversitesi, Muğla

İmza:

Prof. Dr. Cüneyt ÇELİKTAŞ (Üye)
Fizik Anabilim Dalı,
Ege Üniversitesi, İzmir

İmza:

Prof. Dr. Hasan ERTAŞ (Üye)
Kimya Anabilim Dalı,
Ege Üniversitesi, İzmir

İmza:

ANA BİLİM DALI BAŞKANLIĞI ONAYI

Doç. Dr. Çağdaş GÖNEN
Çevre Bilimleri Ana Bilim Dalı Başkanı,
Muğla Sıtkı Koçman Üniversitesi, Muğla

İmza:

Doç. Dr. Oğuz AKPOLAT (Danışman)
Çevre Bilimleri Anabilim Dalı,
Muğla Sıtkı Koçman Üniversitesi, Muğla

İmza:

Dr. Ceyhun KAPUCU (İkinci danışman)
Enformatik Bölüm Başkanlığı,
Muğla Sıtkı Koçman Üniversitesi, Muğla

İmza:

Savunma Tarihi: 16/01/2024

Tez çalışmam boyunca elde ettiğim ve sunduğum tüm bilgi, doküman ve belgelerin akademik ve bilimsel etik kurallarına uygun olduğunu beyan ederim. Ayrıca akademik ve bilimsel etik kurallarına uygun olarak, bu tez çalışması sırasında elde edilmemiş başkalarına ait tüm orijinal bilgilere atıfta bulunduğumu beyan ederim.

Selçuk Kurt
16/01/2024

ÖZET
METEOROLOJİK ÖLÇÜM VERİLERİNİN REGRESYON YAKLAŞIMI İLE
MODELLENMESİNDE VERİ MADENCİLİĞİ AŞAMALARI VE
KULLANIMI

Selçuk KURT

Yüksek Lisans Tezi

Fen Bilimleri Enstitüsü

Çevre Bilimleri Anabilim Dalı

Danışman: Doç. Dr. Oğuz AKPOLAT

İkinci Danışman Dr. Ceyhun KAPUCU

Ocak 2024, 61 sayfa

En yaygın ifadelerle modelleri; doğal benzeşim sistemleri/modelleri, donanımsal (fiziksel) sistemler/modeller ve matematiksel sistemler/modeller olmak üzere üç sınıfa ayırmak mümkündür. Bu çalışmada, meteorolojik ölçüm süreçlerinde sıklıkla ölçülen veriler olarak bağıl nem, hava sıcaklığı, atmosferik basınç, birikmiş günlük toplam yağış miktarı ve ölçümün yapıldığı zaman bilgileri gibi veriler ile güneş ışıınım şiddeti arasındaki ilişkinin modellenmesinde ve güneş ışıınım şiddetinin bazı modeller ile tahmin edilmesi amaçlanmıştır.

Çalışmada ele alınan problem tam anlamıyla bir regresyon problemi olup, çalışmada kullanılan veriler ise literatürde de belirtildiği gibi doğrusal olmayan bir sistemden gelmektedir. Orange veri madenciliği platformunda regresyon problemleri için kullanılabilen modellerden Yapay Sinir Ağları (YSA), Karar Ağacı (KA), Rastgele Orman (RO), Doğrusal Regresyon (DR) ve K-En Yakın Komşu (KEY) modelleri kullanılmıştır. Tüm modellerden elde edilen sonuçlar karşılaştırıldığında regresyon problemine uygun olan ve olmayan modeller belirlenmiştir.

Ayrıca bu tez çalışmasıyla, Orange veri madenciliği platformunun herhangi bir programlama dili bilmeyi gerektirmeden, veri madenciliği, veri modelleme, makine öğrenmesi ile ilgili çalışmalarda kullanılabileceği ve çok hızlı bir şekilde sonuçların elde edilebileceği gösterilmiştir. Çalışmanın bundan sonrası için mimarisinde tıpkı rastgele orman modeli gibi çoklu karar ağacı kullanan gradyan arttırma (Gradient Boosting), kategorik arttırma (CatBoost), uyarlamalı arttırma (AdaBoost) gibi arttırma algoritması temelli modeller ile daha fazla gizli katmanlı YSA modelleri kullanılarak bu tez çalışmasında tercih edilen modellerle karşılaştırılabilir.

Anahtar Kelimeler: Veri Madenciliği, Modelleme, Regresyon, Orange Veri Madenciliği Platformu

ABSTRACT
DATA MINING STAGES AND USAGE IN MODELING
METEOROLOGICAL MEASUREMENT DATA WITH REGRESSION
APPROACH

Selçuk KURT

Master of Science (M.Sc.)

Graduate School of Natural and Applied Sciences

Ecological Sciences

Supervisor: Assoc. Prof. Dr. Oğuz AKPOLAT

Second Advisor Dr. Ceyhun KAPUCU

January 2024, 61 pages

Models in the most common expressions; it is possible to divide them into three classes: natural simulation systems/models, hardware/physical systems/models and mathematical systems/models. In this study, it is aimed to model the relationship between data such as relative humidity, air temperature, atmospheric pressure, accumulated daily total precipitation and the time information when the measurement is made, and solar radiation intensity, which are frequently measured data in meteorological measurement processes, and to estimate solar radiation intensity with some models.

The problem addressed in the study is literally a regression problem, and the data used in the study comes from a nonlinear system, as stated in the literature. Among the models that can be used for regression problems on the Orange data mining platform, Artificial Neural Networks (ANN), Decision Tree (KA), Random Forest (RO), Linear Regression (DR) and K-Nearest Neighbor (KEY) models were used. When the results obtained from all models were compared, models that were and were not suitable for the regression problem were determined.

In addition, with this thesis, it has been shown that the Orange data-mining platform can be used in studies related to data mining data modeling and machine learning and those results can be obtained very quickly, without requiring knowledge of any programming language. For the future of the study, using algorithm-based models such as gradient boosting (Gradient Boosting), categorical boosting (CatBoost), adaptive boosting (AdaBoost) and ANN models with more hidden layers, similar to the random forest model, aimed models for future could be compared with the models preferred in this thesis study.

Keywords: Data Mining, Modelling, Regression, Orange Data Mining Platform

ÖNSÖZ

Tez danışmanlarım Doç. Dr. Oğuz AKPOLAT ve Dr. Ceyhun KAPUCU'ya, çalışmalarım boyunca sağladıkları destek ve rehberlik için içten teşekkürlerimi sunuyorum. Onların bilgi birikimleri ve öğretileri sayesinde, bilimsel düşünce ve çalışma prensiplerini öğrenme fırsatı buldum.

Ayrıca, çalışmamın her aşamasında sabırla yanımda olan sevgili aileme de derin minnettarlığımı ifade etmek isterim.

İÇİNDEKİLER

ÖNSÖZ.....	vii
İÇİNDEKİLER	viii
ÇİZELGELER DİZİNİ	ix
ŞEKİLLER DİZİNİ	x
SEMBOLLER VE KISALTMALAR DİZİNİ	xii
1. GİRİŞ	1
1.1. Meteorolojik Ölçüm Verilerinin Değerlendirilmesi ve Modellenmesinin Önemi4	
1.2. Veri Madenciliği Nedir?.....	4
1.3. Veri Madenciliği Yöntemleri	5
1.4. Veri Madenciliği Yöntemlerinin Meteorolojik Verilerle Kullanılması.....	6
1.5. Veri Madenciliği Yöntemlerinde Kullanılan Aşamalar	8
1.6. Orange Veri Madenciliği Platformu ve Kullanım Alanları.....	9
1.7. Orange Veri Madenciliği Platformunda Kullanılabilen Veri Madenciliği Yöntemleri.....	11
2. MALZEME VE YÖNTEM.....	14
2.1. Deneysel Veri Setinin Elde Edilmesi	14
2.2. Meteorolojik Verilerin Orange’da Hazırlanması.....	17
2.3. Orange’da Regresyon Yaklaşımı ile Veri Madenciliği Uygulaması.....	26
3. BULGULAR VE İRDELEME	33
4. SONUÇLAR VE ÖNERİLER	38
5. KAYNAKLAR	40
6. ÖZGEÇMİŞ	51

ÇİZELGELER DİZİNİ

Çizelge 2.1. Çalışmada kullanılan iklimsel veriler.	21
Çizelge 3.1. Orange veri madenciliği platformunda kullanılan modeller ile elde edilen sonuçlar.....	37



ŞEKİLLER DİZİNİ

Şekil 2.1. Oregon Üniversitesindeki ölçüm istasyonu.	14
Şekil 2.2. Orange platformuna dosyadan veri çekme.	16
Şekil 2.3. Veri çekilen dosyanın özellikleri.	16
Şekil 2.4. Aykırı değerlerin tespit edilmesi.....	19
Şekil 2.5. Normal değerlerle devam edilmesi.	19
Şekil 2.6. Normal değerler listesi.....	20
Şekil 2.7. Aykırı değerler listesi.....	20
Şekil 2.8. Eksik değerlerin doldurulması.	21
Şekil 2.9. Yeni değişkenlerin türetilmesi.	22
Şekil 2.10. Kullanılacak değişkenlerin seçilmesi.....	23
Şekil 2.11. Tüm verilerin normalleştirilmesi.	23
Şekil 2.12. Verilerin gözlenmesi.....	24
Şekil 2.13. Ham veri üzerinde yapılan tüm işlemler.....	25
Şekil 2.14. Değişken istatistikleri ekranı.....	25
Şekil 2.15. Orange'da uçtan uca veri madenciliği süreci.	26
Şekil 2.16. K-En yakın komşu modeli parametreleri.....	28
Şekil 2.17. Doğrusal regresyon modeli parametreleri.....	29
Şekil 2.18. Rastgele orman modeli parametreleri.	30
Şekil 2.19. Yapay sinir ağı modeli parametreleri.....	31
Şekil 2.20. Karar ağacı modeli parametreleri.....	32
Şekil 3.1. K-En yakın komşu modeli sonuçları.	33
Şekil 3.2. Doğrusal regresyon modeli sonuçları.	33

Şekil 3.3. Rastgele orman modeli sonuçları.....	34
Şekil 3.4. Yapay sinir ağı modeli sonuçları.	34
Şekil 3.5. Karar ağacı modeli sonuçları.	35



SEMBOLLER VE KISALTMALAR DİZİNİ

ATA	Arttırma Temelli Algoritma
BM	Bulanık Mantık
DR	Doğrusal Regresyon
GA	Genetik Algoritma
KA	Karar Ağacı
KEY	K-En Yakın Komşu
MAE	Mean Absolute Error – Ortalama Mutlak Hata
MSE	Mean Squared Error – Ortalama Karesel Hata
RMSE	Root Mean Squared Error – Ortalama Karekök Hata
R^2	Coefficient of Determination – Uyum Katsayısının Karesi
YSA	Yapay Sinir Ağları

1. GİRİŞ

Meteorolojik sistemlerdeki çeşitli bileşenler ve parametreler arasındaki ilişkilerin anlaşılması, gerçekliklerin sınıflandırılması ve bu unsurların birbirleriyle olan bağlantılarının keşfedilmesi amacıyla model çalışmaları önem arz etmektedir. Modeller, doğadaki sistemlerin açıklanması, incelenmesi ve analiz edilmesi için kullanılan araçlardır. Model çalışmalarında, öncelik hedeflenen sistem ve oluşturulacak model hakkında fikir, yeterli deneyim, doğru ve güncel bir literatür bilgisi ve ilgili sistemden toplanmış verilere ihtiyaç vardır. Modeller genellikle doğal benzerlik sistemleri, fiziksel (donanımsal) sistemler ve matematiksel sistemler şeklinde üç ana sınıfa ayrılabilir. Bu çalışmada üzerinde durulacak olan modeller ve ilgili uygulamalar, matematiksel sistemler sınıfına dahildir.

Matematiksel bir model oluşturabilmek için, ilgili sistemi anlamak ve tanımlamak amacıyla gerçek deneyler veya benzetimler yoluyla veri toplamak genellikle gereklidir. Bu toplanan veriler, genellikle sayısal büyüklükler olarak temsil edilir ve görsel olarak veri noktaları şeklinde ifade edilebilir. Bu noktalar, sistemdeki serbest/bağımsız değişkenler ile bunlardan etkilenen ve bağımlı olan değişkenleri gösterir. Bu bağımlı değişkenlere göre ifade edilebilecek doğru bir fonksiyon, hedef fonksiyon olarak adlandırılır. Bu fonksiyon, verilerle fiziksel süreç veya sistem arasındaki ilişkiyi tanımlar. Model üzerinden elde edilen sonuçlar, sistemin verilerini anlamlandırabiliyorsa, henüz ölçülmemiş verileri tahmin etmek için de kullanılabilir.

Hedef fonksiyon genellikle $f(x)=y$ şeklinde ifade edilir, burada x bağımsız değişken, y bağımlı değişken ve f hedef fonksiyondur. Ancak, tüm değişkenleri her koşulda hatasız olarak tanımlayan bir hedef fonksiyon belirlemek her zaman mümkün olmayabilir. Bu durum, veri analizi sürecinde sıkça karşılaşılan bir zorluktur. Hedef fonksiyonun ya da hedef fonksiyona en yakın fonksiyonun belirlenmesi, bu sayede modelin hedeflenen sistemdeki örüntüleri tanımlayabilmesi, veri madenciliğinin temel hedeflerinden biridir. Bu süreç, modelin gerçek dünya verilerini anlaması ve gelecekteki olayları tahmin etme yeteneğini artırması açısından kritiktir.

Veri madenciliği teknikleri ile elde edilen modellerin başarısı, bu modellerin gerçek değerlere mümkün olduğunca yakın tahminler yapabilmesine dayanmaktadır. Bu benzerlik derecesini ölçmek için, gerçek değerler (y) ile tahmin edilen değerler (y') arasındaki farkın ölçülmesi yeterlidir. Bu farkın ne kadar küçük olduğu, yani hata ne kadar az ise, tahmin modelinin o kadar başarılı olduğu söylenebilir. Modelin geliştirilme sürecinde, kullanılacak olan parametreler genellikle başlangıçta deneme-yanılma yöntemiyle belirlenir ve daha sonra bu parametreler, yinelemeli algoritmalar içinde güncellenerek her iterasyonda hata kontrol edilir. Bu aşamada, yinelemeli algoritmanın hatanın en az olduğu yöne doğru güncellenmesi amaçlanır. Hatanın en aza indiği nokta bulunduğunda, yineleme durdurulur ve model en uygun parametreleriyle elde edilmiş olur.

Ancak, bu tür yinelemeli algoritmalar genellikle yavaş çalışan, karmaşık hesaplamalar gerektiren ve hata açısından hassas işlemlerdir. Özellikle büyük veri setleri veya gerçek zamanlı uygulamalarda, modelin iyileştirilme sürecini hızlı, güvenilir ve otomatik olarak gerçekleştirebilen, uzman müdahalesini minimize eden tekniklere olan ihtiyaç giderek artmaktadır. Bu, veri madenciliği süreçlerinin etkinliğini artırmak ve model optimizasyonunu daha verimli hale getirmek adına önemli bir hedef haline gelmiştir.

Bu nedenle, karmaşık matematiksel modeller yerine, teknolojinin gelişimiyle birlikte yapay zekâ teknikleri ve araçlarının kullanımı, modelleme, sınıflandırma, regresyon, kümeleme ve tahmin etme gibi alanlarda yaygınlaşmıştır. Bu tür yöntemlerin tercih edilmesindeki en önemli nedenler, matematiksel algoritmaların aksine daha basit olmaları, doğada gerçekleşen birçok doğrusal olmayan sisteme daha uygun olmaları ve öğrenmeye dayalı olmalarıdır. Bu tekniklerin değerlendirilmesinde temel kriterler, basitlik, doğru sonuçlar elde etme yeteneği ve hızdır.

Gelişen bilgisayar teknolojisi ile birlikte, modelleme süreçlerinde yapay zekâ tekniklerinin kullanımı artmaktadır. Bu teknikler arasında Yapay Sinir Ağları (YSA), Karar Ağacı (KA), Rastgele Orman (RO), Doğrusal Regresyon (DR), K-En Yakın Komşu (KEY), Bulanık Mantık (BM), Genetik Algoritma (GA), Arttırma Temelli Algoritmalar (ATA) gibi modeller öne çıkmaktadır. Bu yöntemler, daha karmaşık sistemleri anlamak, öğrenmek ve tahmin etmek için güçlü araçlar sunar ve bu sayede problemlere daha etkili çözümler getirebilir.

Veri analizi ve arařtırmada veri madencilięi aralarını kullanmanın birok faydası bulunmaktadır. zellikle byk veri kmelerinden elde edilen igrlerin verimlilięine, doęruluęuna ve derinlięine katkıda bulunur. Temel faydalardan bazıları řunlardır:

- Karmařık Algoritmaların Verimli Yrtlmesi: Veri madencilięi araları, aęır hesaplamalı veri madencilięi algoritmalarının verimli bir řekilde yrtlmesini saęlayarak, byk ve karmařık veri kmelerinin leklenebilirlik ve gvenilirlikle analizine olanak tanır (Karadimce, 2015).
- rnt Tanımlama ve Tahminler: Veri madencilięi araları, rntlerin tanımlanmasına ve tahminlerin oluřturulmasına olanak tanıyarak karmařık veri kmelerinin anlařılmasına ve analizine katkıda bulunur (Sasikala vd., 2018).
- İstatistiksel Analiz ve Makine ęrenimi: Veri madencilięi araları istatistiksel analiz, makine ęrenimi, rnt tanıma iin gereklidir ve bu alanlardaki ilerlemelere katkıda bulunur (Han, 2011).
- İř Zekası ve Rekabeti Zeka: Veri madencilięi araları, rekabeti zekayı geliřtirmek ve iř kararlarını desteklemek iin kullanılır, iř verilerindeki trendlerin ve kalıpların belirlenmesine katkıda bulunur (Mr, 2021).

zetle, veri madencilięi araları, karmařık algoritmaların verimli bir řekilde yrtlmesi, geliřmiř karar verme sreleri, arařtırmanın kolaylařtırılması ve istatistiksel analiz ve makine ęrenimi desteęi dahil olmak zere ok eřitli faydalar sunabilmektedir.

Bu alıřmada, baęıl nem, hava sıcaklıęı, atmosferik basın, birikmiř gnlk toplam yaęıř miktarı ve lmn yapıldıęı zaman bilgileri gibi meteorolojik lm srelerinde sıklıkla llen veriler ile gneř ıřınım řiddeti arasındaki iliřkinin modellenmesi ve tahmini iin Orange veri madencilięi platformu kullanılacak ve elde edilen sonular paylařılacaktır. alıřmanın sonraki blmnde, bu alanda yapılan literatr alıřmaları ele alınacaktır.

1.1. Meteorolojik Ölçüm Verilerinin Değerlendirilmesi ve Modellenmesinin Önemi

Veri madenciliği, pek çok alanda olduğu gibi meteorolojik veri kümelerinden değerli bilgiler elde etmek için de çok önemli bir rol oynamaktadır. Meteorolojik veri madenciliği, gizli örüntüleri keşfetmek ve büyük ve karmaşık meteorolojik veri kümelerinden yararlı bilgiler çıkarmak için veri madenciliği tekniklerinin uygulanmasını içerir (Zeng vd., 2022). Bu teknikler, büyük ölçekli meteorolojik veri setlerinden örüntüler çıkararak şiddetli fırtına, yoğun yağışlar gibi önemli meteorolojik olayları tahmin etmede özellikle değerlidir (Bartok vd., 2010; Pessoa vd., 2012). Ayrıca veri madenciliği, büyük miktarlardaki meteorolojik veri kümelerinden önceden bilinmeyen örüntülerin otomatik olarak çıkarılmasını sağlayarak meteorolojik olayların anlaşılmasına, analiz edilmesine ve ileriye yönelik tahmin edilmesine katkıda bulunur (Shelke ve Bhagat, 2015; Maimon ve Rokach, 2009).

1.2. Veri Madenciliği Nedir?

Veri madenciliği, mevcut veri kümelerinden, genellikle önceden bilinmeyen modeller biçiminde yeni bilgilerin çıkarılması sürecidir. Büyük veri setlerindeki değerli ve amaca yönelik bilgileri çıkarmak için istatistiksel, matematiksel, yapay zeka ve makine öğrenimi tekniklerinin kullanılmasını içerir. Veri madenciliği özellikle belirli konulara odaklanan araştırmalarda faydalıdır ve sıklıkla daha üst düzey karar alma süreçlerine entegre edilir. Teknoloji, iş dünyası, bilgisayar bilimleri ve çevresel modelleme gibi çeşitli alanlarda yaygın olarak uygulanmış ve bilgi güvenliği, sağlık hizmetleri ve klinik tıp gibi alanlarda değerli bir araç olarak tanımlanmıştır. Bilimsel araştırmalarda veri madenciliği, bilginin keşfedilmesinde ve büyük ve karmaşık veri kümelerinden çıkarılmasında çok önemli bir rol oynar. Yapılandırılmamış, doğal dilden türemiş metinlerden değerli bilgiler elde etmek için günümüzde yaygın olarak kullanılmaktadır. Aynı zamanda zaman serisi verilerindeki potansiyel örüntülerin keşfedilmesi için kullanılmakta ve zaman serisi verilerinden bilginin daha sezgisel bir şekilde çıkarılmasına olanak sağlanmaktadır (Menzies, 2013).

Özetle veri madenciliği, iş, sağlık, çevresel modelleme ve olay analizi dahil olmak üzere çeşitli alanlardaki uygulamalarla büyük veri kümelerinden bilgi keşfi ve

ıkarılması iin gl bir aratır. Arařtırmadaki rol, yapılandırılmamıř metinden bilgi ıkarılması, mahremiyetin korunması, sre denetimleri ve olay verilerinden bilgi ıkarılmasına kadar uzanır.

1.3. Veri Madencilięi Yntemleri

Veri madencilięi yntemleri byk veri kmelerinden deęerli ięrrler ve modeller ıkarmak iin geniř bir yntem yelpazesini kapsar. Bu yntemler eęitim, saęlık hizmetleri, hava tahmini, web veri madencilięi ve daha fazlasını ieren eřitli alanlarda kullanılmaktadır. ne ıkan veri madencilięi yntemlerinden bazıları řunlardır:

- **Sınıflandırma:** Bu yntem, verileri belirli niteliklere gre nceden tanımlanmıř sınıflara veya gruplara ayırmayı ierir. ęrenci performans analizi, saęlık hizmetleri teřhisleri ve web ierięi madencilięi dahil olmak zere eřitli uygulamalarda yaygın olarak kullanılmaktadır (Folorunsho ve Adeyemo, 2012; Fletcher ve İslam, 2014; Patil vd., 2014).
- **Kmeleme:** Kmeleme, benzer veri noktalarının belirli zelliklere gre bir arada gruplanması iřlemidir. Saęlık, tarım ve web ierik madencilięi alanlarında veri kmeleri iindeki kalıpları ve iliřkileri tanımlamak iin kullanılır (Fletcher ve İslam, 2014; Mucherino vd., 2009; Patil vd., 2014).
- **Birliktelik Kuralı Madencilięi:** Bu yntem, byk veri kmelerindeki deęiřkenler arasındaki ilgin iliřkileri keřfetmeye odaklanır. rneęin bir rnn yanında en kolay satılabilecek dięer rnlerin tahmin edilmesi bu yntemin kullanım alanlarındandır. Web verilerinin analizinde, pazarlama alanında, tıbbi teřhislerde ve řıfalı bitkilerle ilgili veri madencilięinde yaygın olarak kullanılır (Patil vd., 2014; Wasan vd., 2006; Babu vd., 2017).
- **Regresyon Analizi:** Bir baęımlı deęiřken ile bir veya daha fazla baęımsız deęiřken arasındaki iliřkiyi anlamak iin regresyon analizi kullanılır. oęunlukla meteorolojik olayları modellemek ve tahmin etmek iin hava tahmini ve iklim deęiřiklięi alıřmalarında kullanılırlar (Kaul, 2017).
- **Anormallik Tespiti:** Anormallik tespit yntemleri, veri kmeleri iindeki aykırı deęerleri veya anormal modelleri tanımlamak iin kullanılır. Olaęandıřı davranıřları veya olayları tespit etmek iin saęlık hizmetleri ve endstriyel veri madencilięinde uygulanırlar (Rahim vd., 2017).

- **Metin Madenciliği:** Metin madenciliği teknikleri, yapılandırılmamış metin verilerinden değerli bilgilerin ve kalıpların çıkarılmasını içerir. Web veri madenciliği ve intihal tespiti gibi alanlarda kullanılırlar (Abid vd., 2017; Kumar, 2017).

Bu veri madenciliği yöntemleri, çeşitli alanlardaki bilgi keşfi, tahmine dayalı modelleme ve karar desteğinde önemli bir rol oynamakta ve bunların büyük veri kümelerinden değerli içgörüler elde etmedeki önemini vurgulamaktadır.

1.4. Veri Madenciliği Yöntemlerinin Meteorolojik Verilerle Kullanılması

Veri madenciliği yöntemlerinin meteorolojik verilerle yapılan uygulamaları şunları içerir:

- Yoğun yağış tahmini için meteorolojik değişkenlerin paterninin belirlenmesi ve analizi (Charles vd., 2017).
- İstatistiksel teknikler ve Monte Carlo simülasyon yöntemleri kullanılarak deniz spreyi¹ buzlanması gibi meteorolojik ve oşinografik koşulların tahmini (Barjouei ve Naseri, 2021).
- Hassas ekim yönetimi için tarım arazisi meteorolojisi tahmin sistemlerinin geliştirilmesi (Ueyama vd., 2018).
- Sayısal hava tahmin modellerinde saatlik yüzey hava sıcaklığı hatalarının düzeltilmesi (Ueyama, 2023).
- Hidrolojik ve meteorolojik veriler kullanılarak topraktaki radyonüklitlerin² bitkilere taşınmasının araştırılması (Pérez-Sánchez ve Thorne, 2014).
- Emisyon verilerinin önemli bir belirsizlik kaynağı olduğu açık ocak madenciliği de dahil olmak üzere atmosferik emisyon envanterlerinin kesinlik ve doğruluğunun değerlendirilmesi (Huertas vd., 2012).
- Hava durumu araştırması ve tahmin modelleri kullanılarak açık ocak madenlerinden kaynaklanan sera gazlarının alan kaçak emisyon akışlarının tahmini (Kia vd., 2022).

¹ Deniz spreyi, okyanus dalgalarının oluşturduğu aerosol haline gelmiş su damlacıklarını ifade eder. Bu küçük parçacıklar rüzgar tarafından taşınabilir ve çeşitli etkilere sahip olabilir.

² Radyonüklit, aşırı nötron veya protonlarla karakterize edilen ve kararsız nükleer enerjiye neden olan çekirdektir.

- Gerçek zamanlı dikey profilli rüzgar verilerini ve hızlı modellemeyi kullanarak karmaşık arazilerde hava akışının tahmin edilmesi (Stocker vd., 2016).
- Karmaşık ortamlarda yer radarları tarafından gözlemlenen meteorolojik bozuklukların analizi ve düzeltilmesi (Wang vd., 2021).
- Güneş enerjisiyle çalışan çok noktalı ağ izlemeyi kullanarak açık ocak madenlerindeki toz konsantrasyonu dağıtım yasasının ve onu etkileyen faktörlerin analizi (Ma vd., 2022).
- Farklı tarım bölgelerinde ampirik referans evapotranspirasyon yöntemlerinin seçimi (Yang vd., 2021).
- Havaalanı rüzgar kesmesi artık döküm uygulamalarında kullanılmak üzere yüksek çözünürlüklü meteorolojik parametrelerin türetilmesi (Stocker vd., 2022).
- Hava durumu araştırması ve tahmin modellerinde yüzey rüzgarı üzerindeki çözümlenmiş ve çözülmemiş topografik etkilerin temsilinin iyileştirilmesi (Jiménez ve Dudhia, 2012).
- En küçük kareler - destek vektör makine modeli aracılığıyla günlük ürün referansı buharlaşma-terleme değerlerinin tahmini (Guo vd., 2011).
- Gradyan artırma tabanlı hiyerarşik zamansal bellek sinir ağları ve bulanık tabanlı sınıflandırma tabanlı regresyon ağacı kullanılarak hava kalitesi yönetiminin iyileştirilmesi (Sagayaraj ve N, 2018).
- Topluluk tahmininde başlangıç ve model belirsizliklerinin etkisini simüle etmek için küresel topluluk tahmin sistemlerinin karşılaştırılması (Buizza vd., 2005).
- Sake pirinci için uygun ekim periyodunu belirlemek için 50 m çözünürlüklü günlük ortalama hava sıcaklığı verilerinin geliştirilmesi (Ueyama vd., 2015).
- Hava tahmin sistemlerinde atmosferik sondaj verilerinin değerlendirilmesi ve özümsemesi (Lu, 2011).
- Farklı kaynaklardan alınan alt stratosfer rüzgâr gözlemlerinin karşılaştırılması (Chen vd., 2022).
- Kapalı madenlerde bir seri tank modeli kullanılarak asit maden drenajı miktarının tahmini (Tokoro vd., 2020).
- Sofar Spotter Ağından spektral dalga verilerinin operasyonel asimilasyonu (Houghton vd., 2022).

- Rüzgârla savrulan toz birikimi tahmini ve maden atıklarının etrafındaki kirliliğin yayılması (Stovern vd., 2016).
- Fotovoltaik santraller için hava kalitesi indeksli sinir ağına dayalı hibrit güç tahmin modeli (Khan vd., 2017).
- Metropol alanlardaki atmosferik kirlleticilerin ve meteorolojik değişkenlerin zaman serilerinin trend birlikteliği analizi (Almanza ve Batyrshin, 2011).

Bu uygulamalar, meteorolojide, hava tahmini ve öngörüsünden çevresel etki değerlendirmelerine ve hava kalitesi yönetimine kadar çeşitli veri madenciliği yöntemlerinin kullanıldığını göstermektedir.

1.5. Veri Madenciliği Yöntemlerinde Kullanılan Aşamalar

Bir veri madenciliği sürecinin aşamaları, büyük veri kümelerinden değerli içgörülerin ve modellerin başarılı bir şekilde çıkarılması için gerekli olan birkaç önemli adımı kapsar. Bu aşamalar veri madenciliği sürecinin doğruluğunu ve güvenilirliğini sağlamak için çok önemlidir. Veri madenciliği sürecinin aşamaları şunları içerir:

- **İlk Keşif:** Bu aşama, verinin yapısını anlamak, ilgili değişkenleri belirlemek ve verinin doğasına dair içgörü kazanmak için veri kümesinin ilk keşfedilme aşamasını içerir (Sasikala vd., 2018).
- **Veri Ön İşleme:** Veri ön işleme, verilerin temizlenmesini, dönüştürülmesini ve analize hazırlanmasını içeren kritik bir aşamadır. Bu, eksik değerlerin ele alınmasını, verilerin normalleştirilmesini ve veri kümesindeki tutarsızlıkların ele alınmasını içerir (Obelets, 2022; Andrews vd., 2019; Eck vd., 2015; Yu vd., 2014; Wegener ve Rüping, 2010; Olaleye ve Tardiff, 2001; Saifuddin ve Ujianto, 2022; Suguna ve Nandhini, 2017; Wegener ve Rüping, 2011; Aulia ve Waspada, 2019; Spruit vd., 2021; Vidyadevi, 2015; Patel ve Patel, 2012; Liu ve Wang, 2007 ; Gandhe vd., 2007; Wang ve Wang, 2008; Chen, 2006; Bilalli vd., 2017; Jabbar vd., 2013; Nugroho ve Prihandoko, 2018; Mahatme ve Bhoyar, 2018; Mariscal vd., 2010 ; Eler vd., 2018; Zulham vd., 2022).
- **Örüntü Belirleme ve Doğrulama:** Bu aşamada, ön işlenmiş verilerdeki örüntüleri, ilişkileri ve eğilimleri belirlemek için veri madenciliği teknikleri uygulanır. Belirlenen örüntüler daha sonra anlamlılıklarını ve güvenilirliklerini

sağlamak için doğrulanır (Sasikala vd., 2018; Obelets, 2022; Andrews vd., 2019; Eck vd., 2015; Yu vd., 2014; Wegener ve Rüping, 2010). ; Olaleye ve Tardiff, 2001; Saifuddin ve Ujjianto, 2022; Suguna ve Nandhini, 2017; Wegener ve Rüping, 2011; Aulia ve Waspada, 2019; Spruit vd., 2021; Vidyadevi, 2015; Patel ve Patel, 2012; Liu ve Wang, 2007; Gandhe vd., 2007; Wang ve Wang, 2008; Chen, 2006; Bilalli vd., 2017; Jabbar vd., 2013; Nugroho ve Prihandoko, 2018; Mahatme ve Bhoyar, 2018; Mariscal vd., 2010; Eler vd., 2018; Zulham vd., 2022).

- **Model Değerlendirmesi:** Örüntü belirlendikten sonra, kullanılan veri madenciliği modelleri, veri seti içindeki temel örüntüleri yakalamadaki performanslarını, doğruluklarını ve etkililiklerini değerlendirmek için değerlendirilir (Obelets, 2022).
- **Model Dağıtım:** Bu aşama, keşfedilen örüntülere dayalı olarak tahminler oluşturmak için veri madenciliği modelinin daha önce hiç görmediği yeni verilerle test edilebilmesi amacıyla dağıtılmasını içerir (Sasikala vd., 2018). Bu dağıtım genellikle elde edilen modelin bir web uygulaması içerisinde bulut bilişim katmanında çalıştırılması şeklinde olabilir.
- **Modellerin Görselleştirilmesi:** Bu aşama, veri seti içerisinde tanımlanan örüntülerin ve ilişkilerin daha iyi anlaşılması için veri madenciliği modellerinin görselleştirilmesini içerir (Obelets, 2022).

Bu aşamalar toplu olarak, büyük ve karmaşık veri kümelerinden değerli içgörüler ve kalıpları ortaya çıkarmak için gerekli olan kapsamlı bir veri madenciliği sürecini oluşturur.

1.6. Orange Veri Madenciliği Platformu ve Kullanım Alanları

Orange, etkileşimli veri analizi, görselleştirme ve modelleme için çeşitli işlevler sunan, makine öğrenimi ve veri madenciliği için kapsamlı bir platformdur. Platform, büyük ve karmaşık veri kümelerinden bilgi keşfi sürecini kolaylaştırmak için tasarlanmıştır. Orange, veri ön işleme, görselleştirme ve modellemeye yönelik bileşenleri bir araya getirerek veri analizi iş akışları oluşturmak için kullanıcı dostu bir ortam sağlar (Demşar vd., 2004; Wang vd., 2009; Podpečan vd., 2011; Arifin vd., 2022; Godec vd., 2019; Sajwan vd., 2023).

Orange veri madenciliği platformunun amaçlanan kullanımları şunları içerir:

- **İnteraktif Veri Analizi:** Orange, verileri keşfetmek ve analiz etmek için etkileşimli bir ortam sunarak kullanıcıların veri seti içindeki yapı ve kalıplara dair içgörü kazanmasına olanak tanır (Demšar vd., 2004).
- **Görselleştirme:** Platform, verileri görsel bir formatta temsil etmek için görselleştirme araçları sağlayarak kullanıcıların karmaşık veri kümelerini daha etkili bir şekilde yorumlayıp anlamasını sağlar (Demšar vd., 2004; Podpečan vd., 2011).
- **Makine Öğrenimi ve Modelleme:** Orange, modelleme ve tahmine dayalı analitik için makine öğrenimi algoritmalarını birleştirerek veri analizi iş akışlarının oluşturulmasını destekler (Demšar vd., 2004; Podpečan vd., 2011; Arifin vd., 2022; Godec vd., 2019).
- **Ön İşleme ve Temizleme:** Orange, veri ön işlemeyi ve temizlemeyi kolaylaştırarak kullanıcıların eksik değerleri ele almasına, verileri normalleştirmesine ve veri kümesi içindeki tutarsızlıkları gidermesine olanak tanır (Demšar vd., 2004; Podpečan vd., 2011).
- **Metin Madenciliği ve Duygu Analizi:** Platform, metin madenciliği ve duygu analizi görevlerini destekleyerek metinsel verilerden anlamlı bilgilerin çıkarılmasına olanak tanır (Sajwan vd., 2023).
- **Eğitimsel Veri Madenciliği:** Orange, eğitim ortamlarından elde edilen verilerden anlamlı bilgiler ortaya çıkarmak için eğitimsel veri madenciliği alanında kullanılmaktadır (Hasan vd., 2020).
- **Bulut Tabanlı Veri Madenciliği:** Platform, bulut tabanlı veri madenciliği faaliyetlerine uygun olup, yüksek performanslı veri yönetimi ve büyük ölçekli veri madenciliği için işlevsellikler sunmaktadır (Cuzzocrea, 2014).

Özetle Orange, eğitim, bulut tabanlı veri madenciliği ve duygu analizi dahil olmak üzere çeşitli alanları kapsayan uygulamalarla etkileşimli veri analizi, görselleştirme, makine öğrenimi ve modelleme için çok yönlü bir platformdur.

1.7. Orange Veri Madenciliği Platformunda Kullanılabilen Veri Madenciliği Yöntemleri

Orange veri madenciliği platformu çok çeşitli veri madenciliği yöntemlerini destekleyerek kullanıcıların çeşitli analitik görevleri gerçekleştirmesine olanak tanır. Orange veri madenciliği platformunda uygulanabilecek veri madenciliği yöntemlerinin bazıları şunlardır:

- **Sınıflandırma:** Orange, sınıflandırma tekniklerini destekleyerek kullanıcıların tahmine dayalı analitik için sınıflandırma modelleri oluşturmaya ve değerlendirmesine olanak tanır (Shamy vd., 2015).
- **Kümeleme:** Platform, kümeleme analizine yönelik işlevler sağlayarak kullanıcıların veriler içindeki doğal gruplamaları belirlemesine olanak tanır (Shamy vd., 2015).
- **Birliktelik Kuralı Madenciliği:** Orange, birliktelik kuralı madenciliği tekniklerini destekleyerek kullanıcıların büyük veri kümelerindeki değişkenler arasındaki ilginç ilişkileri keşfetmesine olanak tanır (Abid vd., 2017).
- **Regresyon Analizi:** Platform, regresyon analizini kolaylaştırarak kullanıcıların değişkenler arasındaki ilişkiyi modellemesine ve verilere dayalı tahminlerde bulunmasına olanak sağlar (Shamy vd., 2015).
- **Metin Madenciliği:** Orange, metin madenciliği yetenekleri sunarak kullanıcıların metinsel verilerden değerli bilgiler çıkarmasına ve duygu analizi yapmasına olanak tanır (Abid vd., 2017; Habala vd., 2012).
- **Eğitimsel Veri Madenciliği:** Platform, eğitimsel veri madenciliği tekniklerini destekleyerek öğrencilerin eğitim ortamlarındaki performanslarının ve etkileşimlerinin analizine olanak sağlar (Sharma, 2020; Baradwaj ve Pal, 2011; He, 2013).
- **Veri Ön İşleme:** Orange, eksik değerlerin ele alınması, normalleştirme ve verilerin temizlenmesi dahil olmak üzere veri ön işleme için işlevler sunar (Shamy vd., 2015).
- **Görselleştirme:** Platform, verileri ve analiz sonuçlarını görsel bir formatta temsil etmek için görselleştirme araçları sağlar ve verilerin yorumlanmasına ve anlaşılmasına yardımcı olur (Shamy vd., 2015).

Orange veri madenciliği platformu, birçok veri madenciliği ve makine öğrenimi algoritmasını destekler, bu nedenle kullanıcılar geniş bir araç yelpazesıyla çeşitli analitik görevleri gerçekleştirebilir. Orange platformunda uygulanabilen bazı veri madenciliği modelleri;

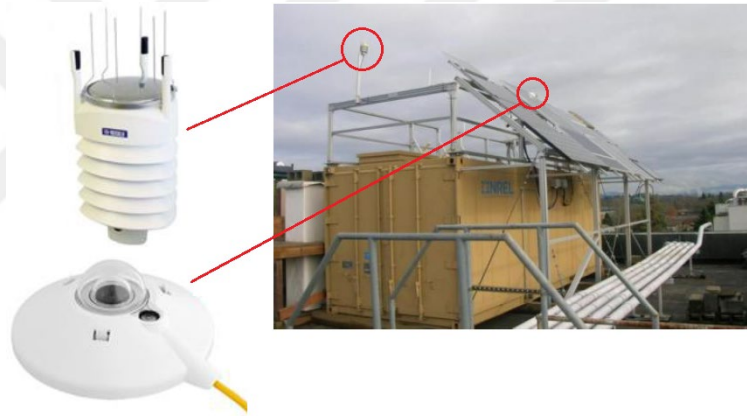
- **Karar Ağacı (Decision Tree):** Orange, karar ağaçları oluşturarak sınıflandırma ve regresyon analizleri yapmanıza olanak tanır. Karar ağaçları, veri kümesindeki ilişkileri anlamak ve görselleştirmek için etkili bir yöntemdir.
- **Yapay Sinir Ağları (Neural Networks):** Yapay sinir ağları, karmaşık yapıdaki veri setlerinde öğrenme yeteneği olan modeller oluşturabilir. Orange, yapay sinir ağlarını oluşturmak ve eğitmek için araçlar sunar.
- **Doğrusal Regresyon (Linear Regression):** Orange, doğrusal regresyon modelleri oluşturarak bir değişkenin diğer değişken üzerindeki etkisini anlamak için kullanılabilir.
- **Destek Vektör Makineleri (Support Vector Machines):** Orange, sınıflandırma ve regresyon problemlerini çözmek için destek vektör makineleri (SVM) kullanmanıza olanak tanır.
- **K-Means Kümeleme (K-Means Clustering):** Kümeleme analizi için Orange, K-Means algoritmasını kullanarak veri noktalarını belirli kümelerde gruplamak için bir araç sağlar.
- **Rastgele Orman (Random Forest):** Rastgele ormanlar, sınıflandırma ve regresyon görevlerinde kullanılan etkili bir ensemble öğrenme yöntemidir. Orange, rastgele orman modelleri oluşturmak için bir araç sunar.
- **Bulanık Mantık (Fuzzy Logic):** Bulanık mantık, belirsiz veri kümeleri üzerinde analiz yapmak için kullanılır. Orange platformu, bulanık mantık modelleri oluşturmak için bir modül sunar.
- **Genetik Algoritmalar (Genetic Algorithms):** Genetik algoritmalar, evrimsel optimizasyon problemlerini çözmek için kullanılır. Orange, genetik algoritma kullanarak model parametrelerini optimize etmek için bir araç sağlar.
- **K-En Yakın Komşu (K-Nearest Neighbors):** Orange, K-en yakın komşu algoritmasını kullanarak sınıflandırma ve regresyon problemlerini çözmek için bir araç sunar.
- **PCA (Principal Component Analysis):** Orange, PCA kullanarak boyut azaltma ve veri setlerindeki varyansı anlamak için bir modül sağlar.

Özetle, Orange veri madenciliği platformu, onu çeşitli alanlardaki çok çeşitli analitik görevlere uygun hale getiren kapsamlı bir veri madenciliği teknikleri seti sunar. Platform, kullanıcıların bu yöntemleri veri setlerine uygulayarak çeşitli analitik görevleri gerçekleştirmelerine imkân tanır.



2. MALZEME VE YÖNTEM

Tez çalışmasında, Eugene, Oregon'daki Oregon Üniversitesi'nde Amerikan Ulusal Yenilenebilir Enerji Laboratuvarı (NREL) tarafından kurulan ölçüm istasyonundan alınan veriler kullanılmıştır (Marion vd., 2014; “Photovoltaic Research, NREL,” 2019). Bu veriler, güneş enerjisini elektrik enerjisine dönüştüren güneş panellerinin performansını modelleme amacıyla kaydedilmiş geniş çaplı bir veri setinden elde edilmiştir. Veriler, halka açık bir kaynaktan erişilebilir olup, kuruma yapılan talep doğrultusunda sağlanan dosya indirme adresinden indirilmiştir. Bu ölçüm verileri, güneş panellerinin çeşitli iklim koşullarında nasıl tepki verdiğini anlamak ve güneş enerjisi sistemlerinin performansını optimize etmek için önemli bir kaynak oluşturmaktadır.

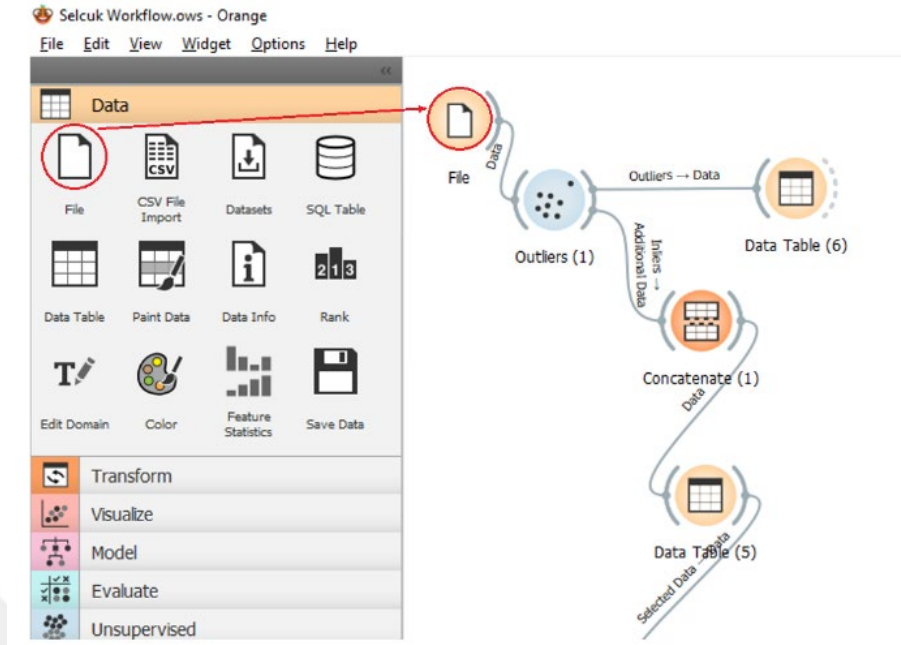


Şekil 2.1. Oregon Üniversitesi'ndeki ölçüm istasyonu.

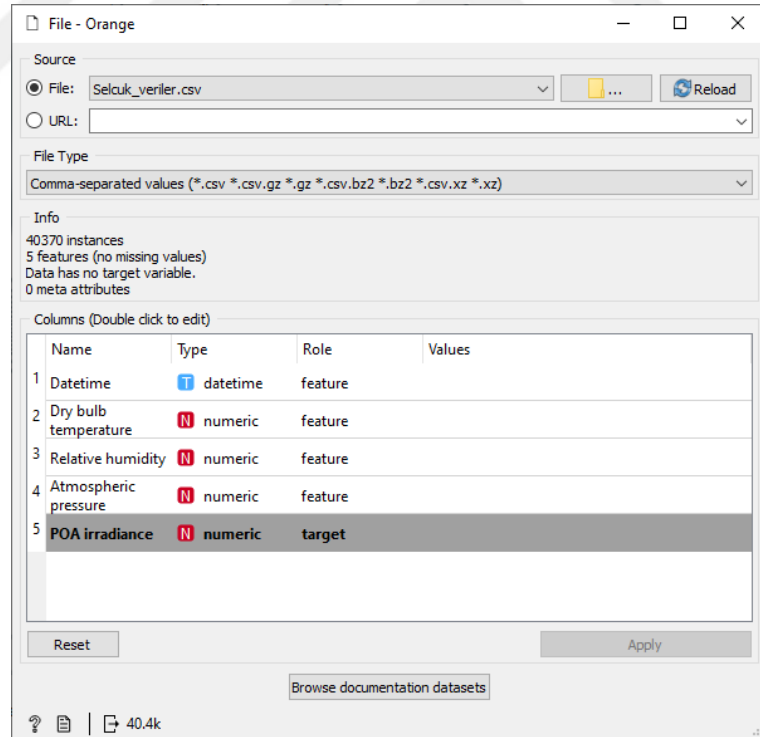
Bir sonraki bölümde ise ham veri setinden deneysel veri setinin nasıl elde edildiği ele alınmıştır.

2.1. Deneysel Veri Setinin Elde Edilmesi

Tez çalışmasında ele alınan ham veri seti, iklimsel ve elektriksel ölçüm verilerini içermektedir. Ancak, tezin temel odak noktası güneş ışıyım şiddetinin modellenmesi olduğundan, elektriksel veriler kullanılmamıştır. Ana amaç, belirli iklimsel değişkenlerin, özellikle güneş ışıyım şiddetinin modellenmesidir. Kullanılan iklimsel



Şekil 2.2. Orange platformuna dosyadan veri çekme.



Şekil 2.3. Veri çekilen dosyanın özellikleri.

Bu ayar ekranında veri setindeki kayıt sayısı, değişken sayısı, değişkenlerdeki eksik değer sayısı gibi bilgiler de verilmektedir. Hedef değişkenin hangi değişken olacağı

kullanıcı tarafından bu ekranda seçilmektedir. Bu ekranda yerel bir dosya haricinde internetteki bir adreste yer alan dosya da veri kaynağı olarak gösterilebilir.

Ölçümün yapıldığı zaman bilgisi, *yıl – ay – gün saat: dakika: saniye* cinsinden kayıtlı olduğundan, zaman damgası türünde bir değişkendir bu değişken kullanılarak yeni değişkenler türetilmiştir. Bu amaçla yapılan işlemler bir sonraki bölümde ele alınmıştır.

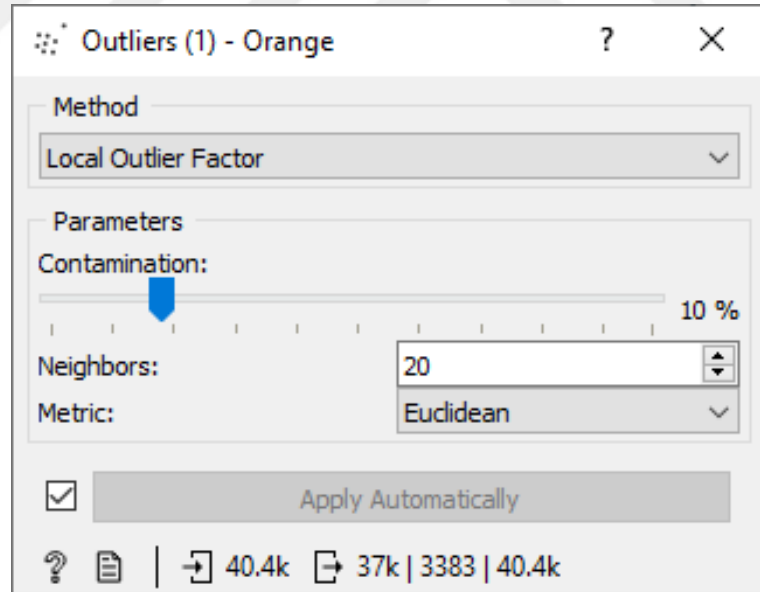
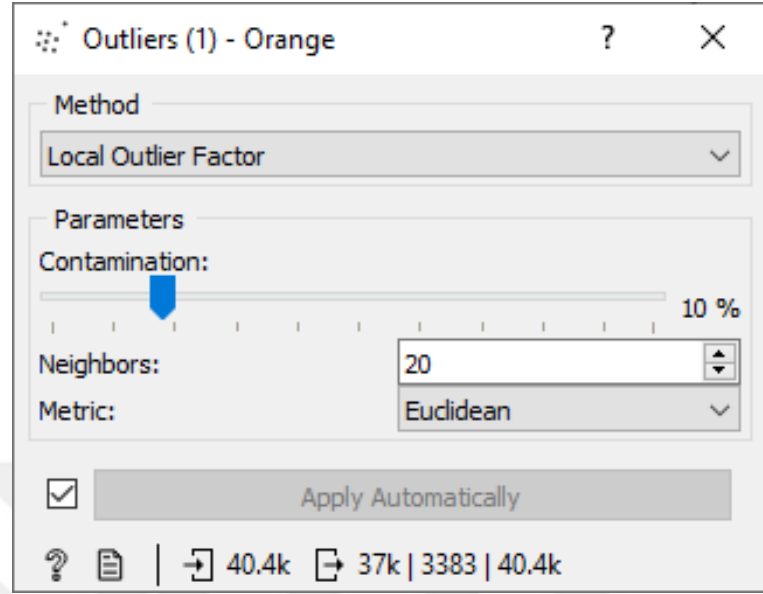
2.2. Meteorolojik Verilerin Orange’da Hazırlanması

Tez çalışmasında sahip olunan ham verilere uygulanan veri temizleme ve değişken mühendisliği adımları şu şekildedir.

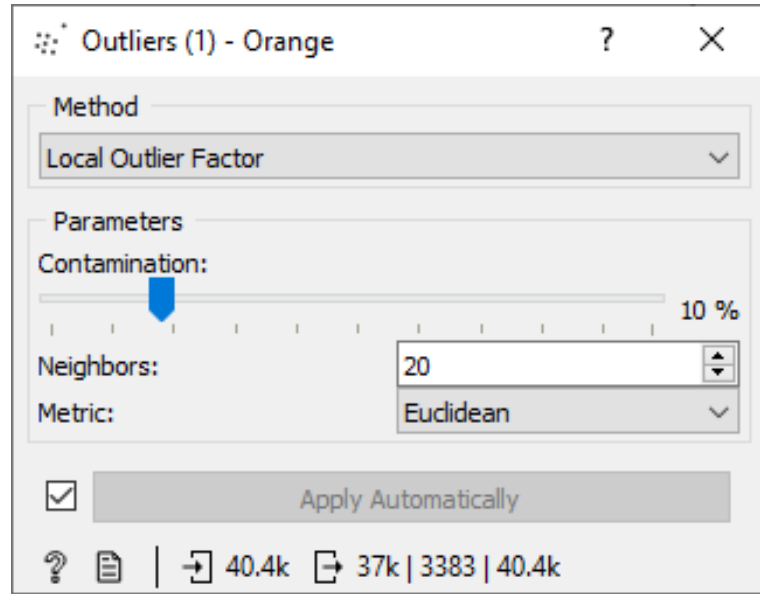
- Aykırı değer temizleme
- Eksik değer doldurma
- Yeni değişkenlerin oluşturulması
- Gereksiz değişkenlerin çıkarılması

Aykırı değerler istatistiksel bir analizde, diğer veri noktalarından belirgin şekilde farklı olan değerleri ifade eder. Bu değerler, genellikle ölçüm hatası veya popülasyonun genel dağılımından sapmalardan kaynaklanabilir. Orange’da aykırı değer tespiti için literatürde de yaygın olarak kullanılan LOF (Local Outlier Factor – Yerel Aykırı Değer Faktörü) yöntemi kullanılmıştır. Yerel aykırı değer faktörü, veri biliminde kullanılan bir kavramdır ve aykırı değerleri tespit etmeye yönelik bir metriği ifade eder. Bu faktör, belirli bir veri noktasının, çevresindeki diğer noktalardan ne kadar farklı olduğunu ölçer. Yüksek bir yerel aykırı değer faktörü, bir noktanın aykırı olduğunu gösterirken, düşük bir değer normal davranışı yansıtabilir. Bu farkı ölçerken Öklit, Manhattan vb mesafe ölçütleri kullanılır.

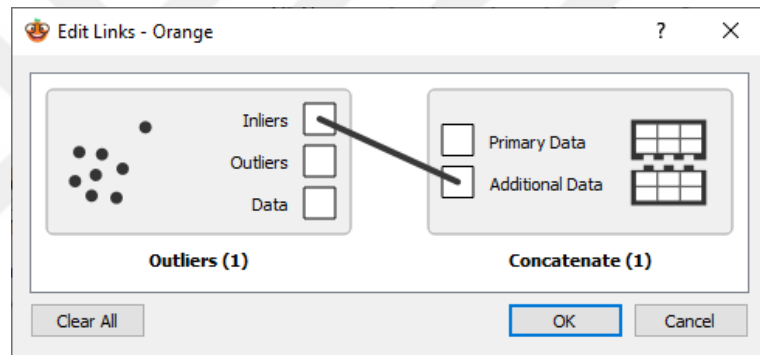
Çalışmada Orange’da bulunan “Outliers” aracı kullanılarak normal ve aykırı değerler tespit edilmiş ve çalışmaya normal değerler ile devam edilmiştir.



Şekil 2.4, Şekil 2.5).



Şekil 2.4. Aykırı değerlerin tespit edilmesi.



Şekil 2.5. Normal değerlerle devam edilmesi.

Ham verisetindeki örnek sayısı en başta 40370 iken, aykırı değerlerin elenmesinden sonra bu sayı 36987'ye düşmüştür (Şekil 2.6). Elenen aykırı değer sayısı 3383'tür (Şekil 2.7).

Data Table (5) - Orange

Info
36987 instances (no missing data)
4 features
Numeric outcome
No meta attributes.

Variables
☒ Show variable labels (if present)
☐ Visualize numeric values
☒ Color by instance classes

Selection
☒ Select full rows

Restore Original Order

☒ Send Automatically

37k | 37k

	POA irradiance	Datetime	ry bulb temperatur	Relative humidity	tmospheric pressur
36967	137.4	2013-12-31 14:2...	11.3	59.2	1006.3
36968	120.0	2013-12-31 14:3...	11.4	59.1	1006.2
36969	112.2	2013-12-31 14:3...	11.4	58.2	1006.2
36970	108.2	2013-12-31 14:4...	11.5	56.1	1006.3
36971	85.3	2013-12-31 14:4...	11.5	56.5	1006.3
36972	72.9	2013-12-31 14:5...	11.5	56.2	1006.3
36973	77.8	2013-12-31 14:5...	11.5	55.1	1006.5
36974	71.4	2013-12-31 15:0...	11.6	55.5	1006.5
36975	63.5	2013-12-31 15:0...	11.7	53.7	1006.5
36976	100.3	2013-12-31 15:1...	11.7	53.3	1006.5
36977	130.3	2013-12-31 15:2...	11.7	53.4	1006.6
36978	448.1	2013-12-31 15:2...	11.7	53.4	1006.6
36979	413.4	2013-12-31 15:3...	12.0	52.8	1006.8
36980	381.9	2013-12-31 15:3...	12.1	52.6	1006.7
36981	360.6	2013-12-31 15:4...	12.0	52.4	1006.7
36982	343.8	2013-12-31 15:4...	12.2	51.4	1006.7
36983	312.9	2013-12-31 15:5...	12.1	51.4	1006.8
36984	299.9	2013-12-31 15:5...	12.1	51.7	1006.7
36985	282.8	2013-12-31 16:0...	12.3	51.5	1006.8
36986	262.0	2013-12-31 16:0...	12.2	50.8	1006.8
36987	209.7	2013-12-31 16:1...	12.2	51.0	1006.9

Şekil 2.6. Normal değerler listesi.

Data Table (6) - Orange

Info
3383 instances (no missing data)
4 features
Numeric outcome
No meta attributes.

Variables
☒ Show variable labels (if present)
☐ Visualize numeric values
☒ Color by instance classes

Selection
☒ Select full rows

Restore Original Order

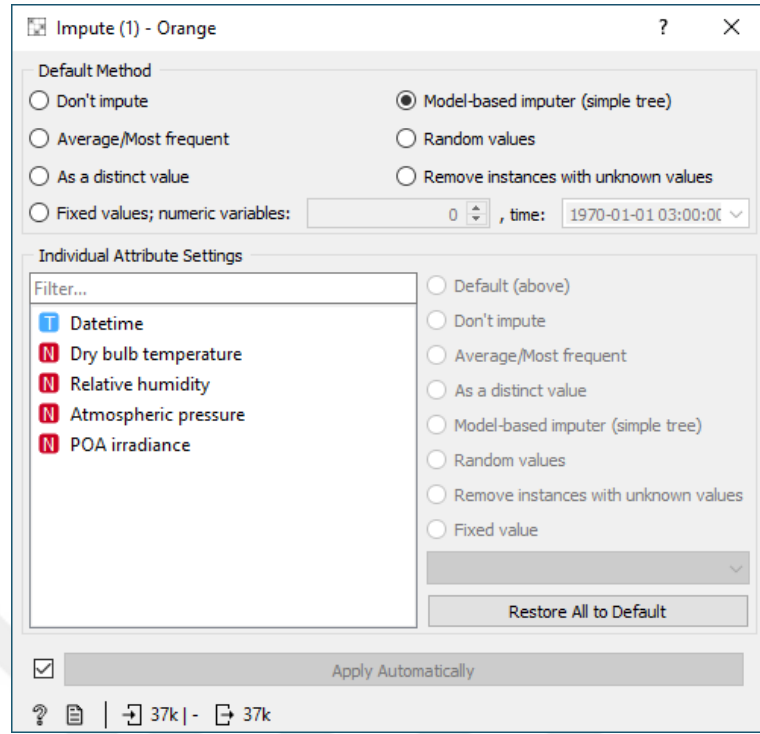
☒ Send Automatically

3383 | 3383

	POA irradiance	Datetime	ry bulb temperatur	Relative humidity	tmospheric pressur
3363	29.5	2013-12-29 15:4...	3.7	81.7	1008.2
3364	26.4	2013-12-29 15:4...	3.8	82.0	1008.1
3365	25.4	2013-12-29 15:5...	3.8	81.7	1008.0
3366	22.8	2013-12-29 15:5...	3.8	81.6	1007.9
3367	21.5	2013-12-30 08:4...	2.8	84.9	1004.5
3368	28.7	2013-12-30 08:4...	2.8	84.5	1004.4
3369	34.9	2013-12-30 08:5...	2.8	83.4	1004.3
3370	46.4	2013-12-30 08:5...	2.9	83.9	1004.4
3371	63.0	2013-12-30 09:0...	3.0	83.2	1004.4
3372	47.7	2013-12-30 16:0...	7.4	75.9	1005.4
3373	44.1	2013-12-30 16:0...	7.5	75.8	1005.4
3374	39.2	2013-12-30 16:1...	7.4	76.0	1005.4
3375	49.0	2013-12-30 16:1...	7.4	76.1	1005.4
3376	38.6	2013-12-30 16:2...	7.3	76.8	1005.3
3377	21.7	2013-12-31 08:3...	4.3	85.0	1007.0
3378	25.6	2013-12-31 08:4...	4.3	85.2	1007.0
3379	29.5	2013-12-31 08:4...	4.7	84.3	1006.9
3380	65.0	2013-12-31 16:1...	12.2	51.1	1006.8
3381	38.1	2013-12-31 16:2...	12.4	51.3	1006.8
3382	24.5	2013-12-31 16:2...	12.4	52.2	1006.8
3383	24.3	2013-12-31 16:3...	12.1	54.8	1006.8

Şekil 2.7. Aykırı değerler listesi.

Veri setindeki eksik değerlerin doldurulması için Orange’da yer alan “Impute” aracı kullanılmıştır (Şekil 2.8). Veri setinde herhangi bir eksik değer olmadığından bu araç çalışmamış fakat uçtan uca bir veri madenciliği uygulamasında gerekli bir adım olduğundan yine de sürece eklenmiştir.



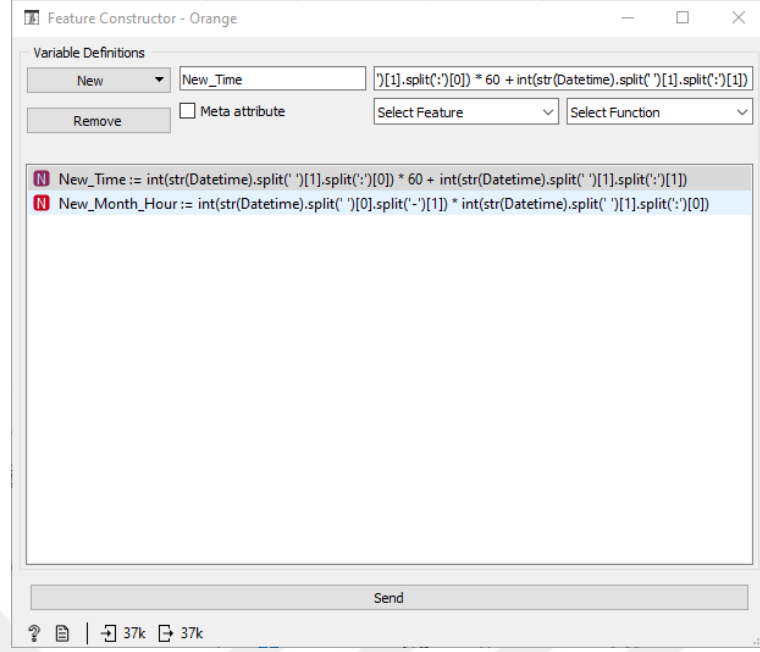
Şekil 2.8. Eksik değerlerin doldurulması.

Bu adımlardan sonra elde edilen iklimsel veriler hakkında istatistiki bilgiler Çizelge 2.1’de listelenmiştir.

Çizelge 2.1. Çalışmada kullanılan iklimsel veriler.

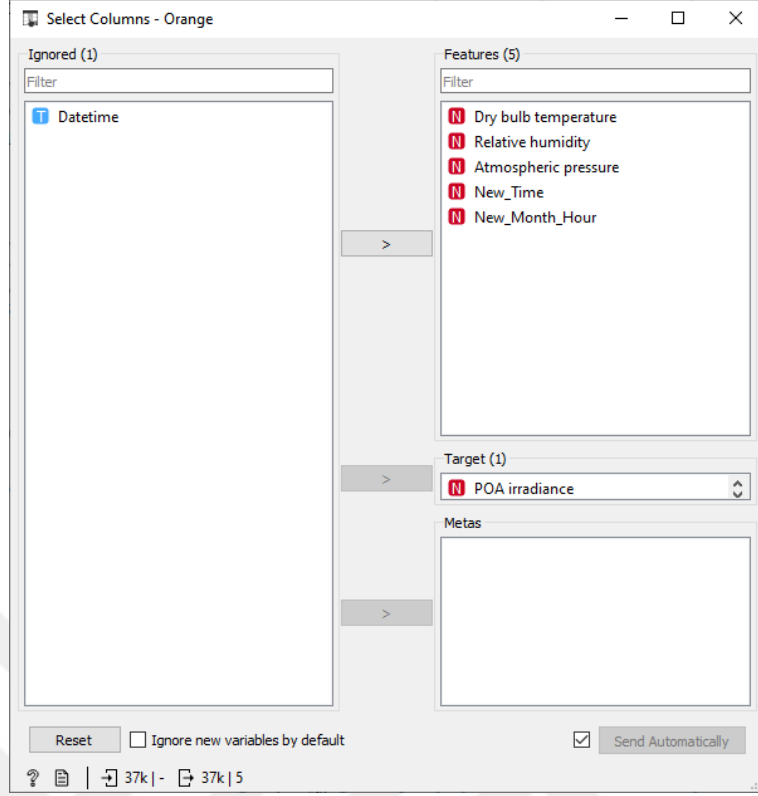
Veri tipi	En küçük	En büyük	Ortalama
Güneş ışıınım şiddeti	20,1W/m ²	1377,9W/m ²	397,66W/m ²
Hava sıcaklığı	-7,4°C	33,7°C	15,16°C
Bağıl nem	%13,9	%93,9	%61,98
Atmosferik basınç	984,1mb	1017,7mb	1000,37mb

Bir sonraki adım olarak veri setindeki zaman bilgisinden yeni değişkenler türetilmiş ve bunun için Orange’da yer alan “Feature Constructor” aracı kullanılmıştır (Şekil 2.9). Bu araç ekranında, matematiksel dönüşüm ve işlemler kullanılarak zaman bilgisinden zaman içindeki saat ve ay-saat çarpımı gibi iki yeni değişken türetilmiştir.



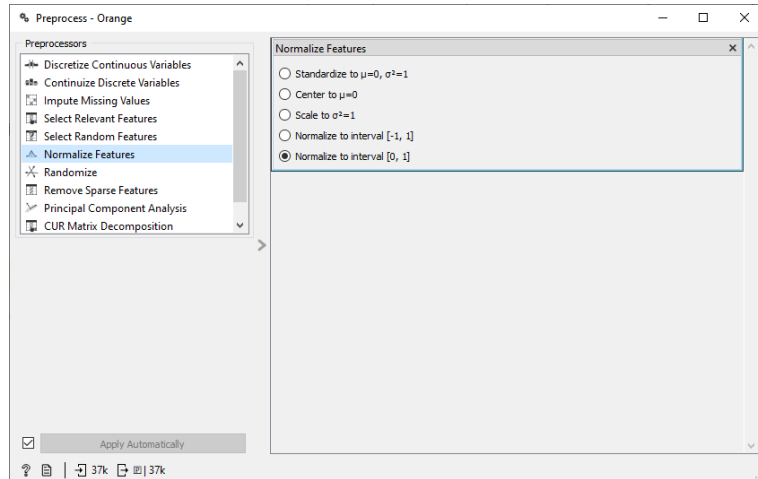
Şekil 2.9. Yeni değişkenlerin türetilmesi.

Türetilen bu yeni değişkenlerin ardından analiz sırasında kullanılmayacak olan zaman bilgisi verisetinden çıkarılmış ve sürece diğer değişkenlerle devam edilmiştir. Bu değişken seçimi için Orange’da yer alan “Select Columns” aracı kullanılmıştır (Şekil 2.10).



Şekil 2.10. Kullanılacak değişkenlerin seçilmesi.

Süreçte kullanılacak değişkenler seçildikten sonra tüm değişkenlerin aynı aralıkta [0, 1] olmaları için normalleştirme yapılmıştır. Bunun için Orange’da yer alan “Preprocess” aracı kullanılmıştır (Şekil 2.11).

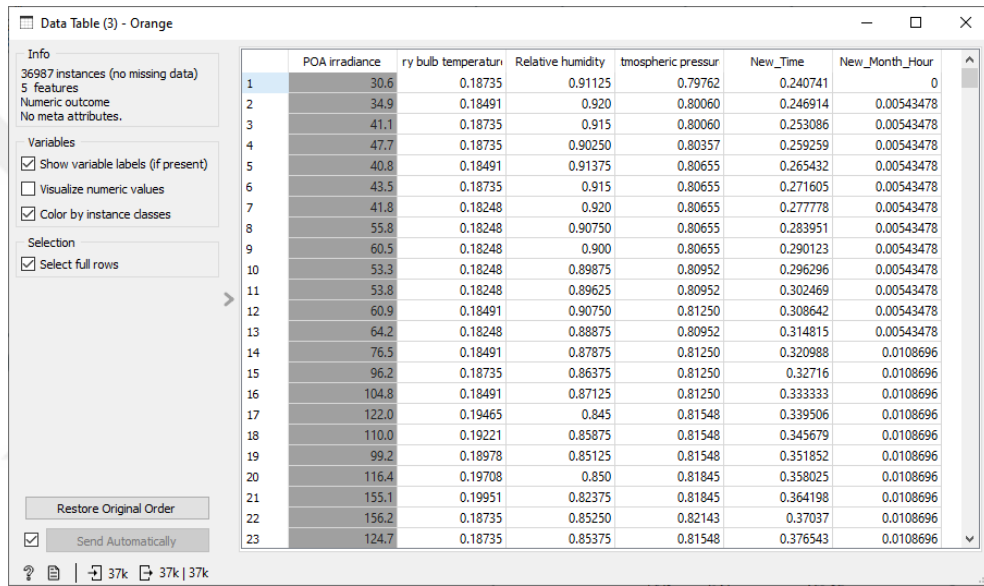


Şekil 2.11. Tüm verilerin normalleştirilmesi.

Veri madenciliği uygulamalarında verilerin normalleştirilmesinin veri analizi üzerinde önemli etkileri vardır. Gürültüyü bastırmaya, performansı artırmaya, doğruluğu artırmaya, madencilik için gereken süreyi en aza indirmeye yardımcı olur (Yang ve

Gao, 2010; Ismail vd., 2015; Liu vd., 2016). Ek olarak verilerden anlamlı ilişkilerin çıkarılmasına olanak tanıyarak örüntülerin ortaya çıkarılmasında ve operasyonların optimize edilmesinde daha etkili olmasını sağlar (Zanev vd., 2014). Bu nedenle veri normalizasyonu, veri madenciliği süreçlerinin etkinliğini ve verimliliğini arttırmada çok önemli bir rol oynamaktadır.

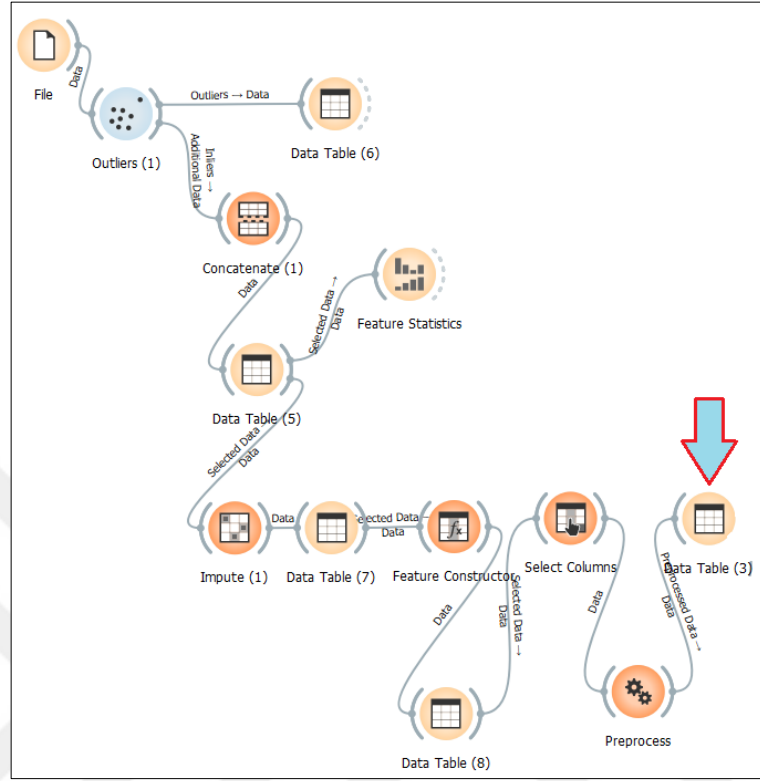
Tüm bu adımların ardından veriler Orange’da modellerin eğitilmesi için kullanıma hazır hale gelmiştir. Verilerin Orange’da veri madenciliği sürecinin herhangi bir yerinde gözlenmesi, kontrol edilmesi mümkündür. Bu amaçla Orange’da yer alan “Data Table” aracı süreç içerisinde pek çok yerde kullanılmıştır (Şekil 2.12).



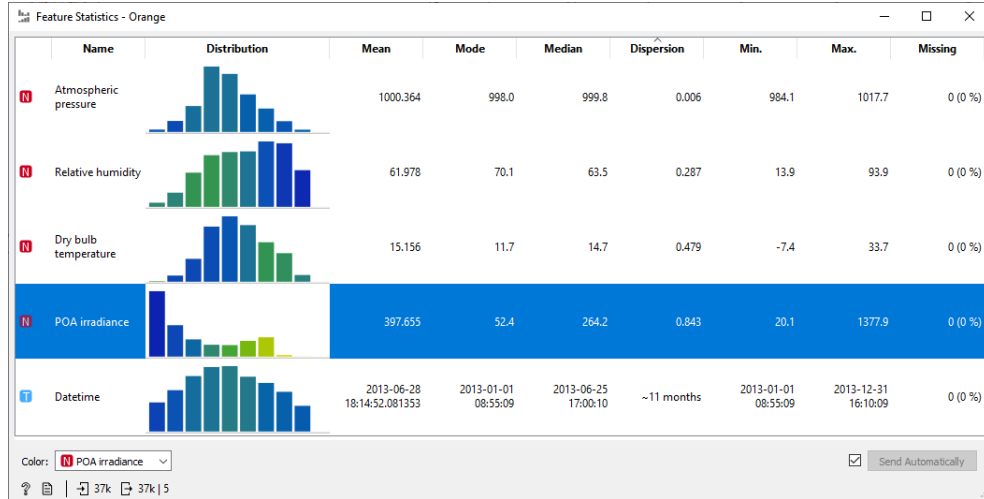
	POA irradiance	ry bulb temperatur	Relative humidity	tmospheric pressur	New_Time	New_Month_Hour
1	30.6	0.18735	0.91125	0.79762	0.240741	0
2	34.9	0.18491	0.920	0.80060	0.246914	0.00543478
3	41.1	0.18735	0.915	0.80060	0.253086	0.00543478
4	47.7	0.18735	0.90250	0.80357	0.259259	0.00543478
5	40.8	0.18491	0.91375	0.80655	0.265432	0.00543478
6	43.5	0.18735	0.915	0.80655	0.271605	0.00543478
7	41.8	0.18248	0.920	0.80655	0.277778	0.00543478
8	55.8	0.18248	0.90750	0.80655	0.283951	0.00543478
9	60.5	0.18248	0.900	0.80655	0.290123	0.00543478
10	53.3	0.18248	0.89875	0.80952	0.296296	0.00543478
11	53.8	0.18248	0.89625	0.80952	0.302469	0.00543478
12	60.9	0.18491	0.90750	0.81250	0.308642	0.00543478
13	64.2	0.18248	0.88875	0.80952	0.314815	0.00543478
14	76.5	0.18491	0.87875	0.81250	0.320988	0.0108696
15	96.2	0.18735	0.86375	0.81250	0.32716	0.0108696
16	104.8	0.18491	0.87125	0.81250	0.333333	0.0108696
17	122.0	0.19465	0.845	0.81548	0.339506	0.0108696
18	110.0	0.19221	0.85875	0.81548	0.345679	0.0108696
19	99.2	0.18978	0.85125	0.81548	0.351852	0.0108696
20	116.4	0.19708	0.850	0.81845	0.358025	0.0108696
21	155.1	0.19951	0.82375	0.81845	0.364198	0.0108696
22	156.2	0.18735	0.85250	0.82143	0.37037	0.0108696
23	124.7	0.18735	0.85375	0.81548	0.376543	0.0108696

Şekil 2.12. Verilerin gözlenmesi.

Veri madenciliği sürecinin en başından itibaren sahip olunan ham veriye uygulanan veri temizleme ve değişken mühendisliği dahil tüm işlemler Orange çalışma ekranında aşağıdaki iş akışı ile gerçekleştirilmiştir (Şekil 2.13). Bu iş akışında modelleme, modellerin değerlendirilmesi, görselleştirilmesi gibi adımlar gösterilmemiştir. Şekil 2.12’deki veri listesi görünümü, Şekil 2.13’de ki renkli okla işaretlenmiş “Data Table” üzerinden elde edilmiştir. Şekil 2.13’de görüldüğü gibi sürecin bazı aşamalarında verileri gözlemlemek ve kontrol etmek için pekçok “Data Table” aracı kullanılmıştır. Şekil 2.13’de kullanıldığı görülen “Feature Statistics” aracı, sahip olunan tüm değişkenler için en küçük, en büyük, ortalama ve medyan değerleri gibi istatistiksel ölçümleri ve değişkenin dağılım grafiğini görmek için kullanılmıştır (Şekil 2.14). Ek olarak Çizelge 2.1’de listelenen bilgiler bu araç kullanılarak elde edilmiştir.



Şekil 2.13. Ham veri üzerinde yapılan tüm işlemler.

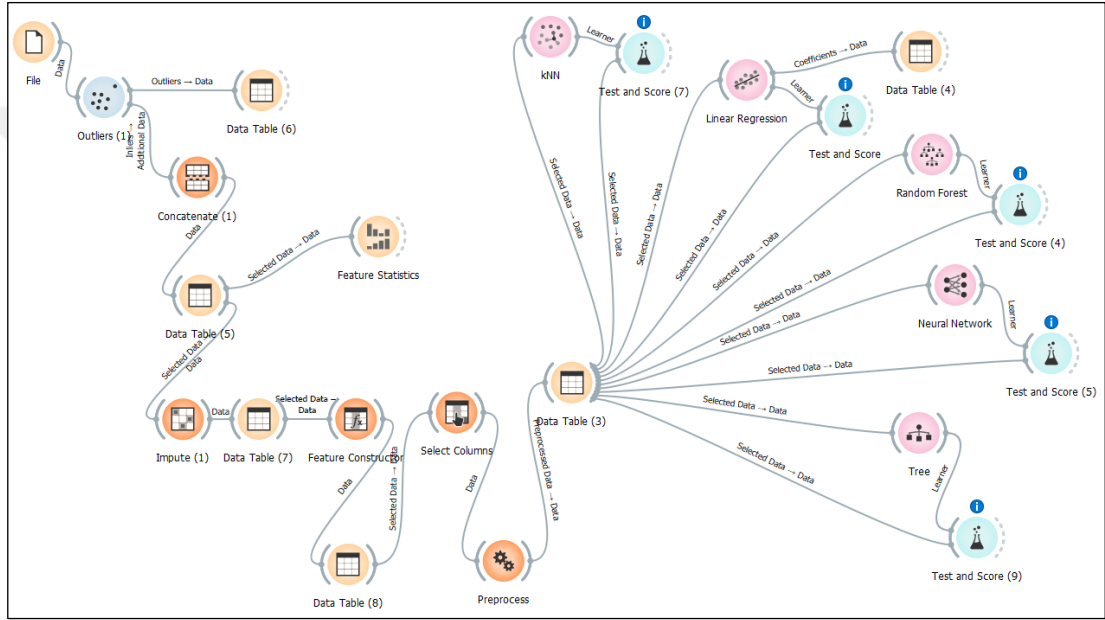


Şekil 2.14. Değişken istatistikleri ekranı.

Bir sonraki bölümde elde edilen veriseti kullanılarak meteorolojik verilerin modellenmesi ele alınacaktır.

2.3. Orange’da Regresyon Yaklaşımı ile Veri Madenciliği Uygulaması

Çalışmada kullanılacak veri madenciliği modelleri olarak Yapay Sinir Ağları (YSA), Karar Ağacı (KA), Rastgele Orman (RO), Doğrusal Regresyon (DR) ve K-En Yakın Komşu (KEY) modelleri kullanılmıştır. Baştan sona tüm veri madenciliği süreci Orange’da oluşturulmuştur (Şekil 2.15). Bu süreç Orange’da iş akışı (Orange Workflow) adıyla anılır ve ows dosya uzantısıyla kalıcı olarak saklanabilir. Bu dosya istenirse paylaşılarak veri madenciliği sürecinin araştırmacıların kullanımına açılması sağlanabilir.



Şekil 2.15. Orange’da uçtan uca veri madenciliği süreci.

Şekil 2.15’e göre kullanılacak modeller veri madenciliği sürecine eklenmiş ve her modelin çıkışına modeli test etmek ve tahminleme başarısını görmek için “Test and Score” aracı eklenmiştir. Bu araç ile hedeflenen veri madenciliği modelinin, sahip olunan verilerle nasıl eğitileceği, verinin ne kadarının eğitim ne kadarının test amacıyla kullanılacağı gibi ayarlamalar yapıldıktan sonra eğitilmiş modelin test verilerinde elde ettiği tahmin performansları görülebilir.

Tez çalışmasında kullanılan Orange platformu, kullanılan modellerin tahmin performansını göstermek amacıyla; ortalama karesel hata (Mean Squared Error – MSE), ortalama karekök hata (Root Mean Squared Error – RMSE), ortalama mutlak hata (Mean Absolute Error – MAE) ve uyum katsayısının karesi (R^2) ölçütlerini bir arada kullanmaktadır. Bunlar oldukça yaygın kullanılan performans ölçütleridir.

MSE ölçütünü tanımlayan eşitlik aşağıda verilmiştir. Burada t değişkeni gerçek güneş ışıınımı verileri, y değişkeni model tarafından tahmin edilen güneş ışıınımı verilerini gösterir.

$$MSE = \frac{1}{n} \sum_{i=1}^n (t_i - y_i)^2 \quad (1)$$

RMSE ölçütünü tanımlayan eşitlik aşağıda verilmiştir.

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (t_i - y_i)^2} \quad (2)$$

MAE ölçütünü tanımlayan eşitlik aşağıda verilmiştir.

$$MAE = \frac{1}{n} \sum_{i=1}^n |t_i - y_i| \quad (3)$$

R^2 ölçütünü tanımlayan eşitlik aşağıda verilmiştir. Burada t_o değişkeni ise gerçek güneş ışıınımı verilerinin ortalamasıdır.

$$R^2 = 1 - \frac{\sum_{i=1}^n (t_i - y_i)^2}{\sum_{i=1}^n (t_i - t_o)^2} \quad (4)$$

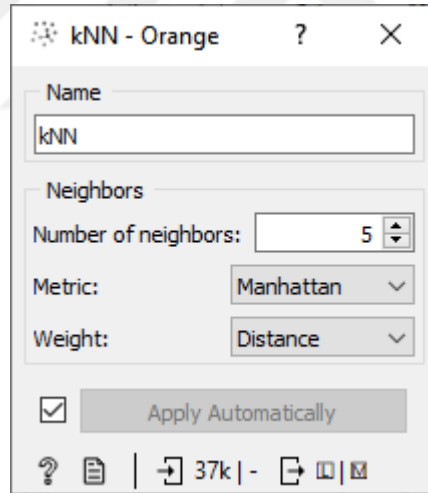
MSE, RMSE ve MAE değerlerinin olabildiğince küçük olması modelin tahmin başarısının yüksek olduğu anlamına gelmektedir. RMSE, tahmin edilen değer ile gerçek değer arasındaki farkın büyük olduğu durumlarda performansı yansıtmaya açısından daha çok tercih edilmektedir. R^2 değeri ise 0 ile 1 arasında değişmektedir. Bu değer 1'e yaklaşması modelin tahmin ettiği güneş ışıınımı değerlerinin, gerçek değerlerle örtüştüğü anlamına gelmektedir. R^2 ölçütünün birden fazla modelin tahminleme başarısını karşılaştırırken daha anlaşılır bir ölçüt olduğunu söyleyebiliriz. Örneğin R^2 değeri 0.8 olan bir model 0.7 olan bir modelden daha başarılıdır denilebilir.

Tez çalışması kapsamında oluşturulan tüm modellerin test edilebilmesi için 5 katlı çapraz doğrulama kullanılmıştır. K-Katlı çapraz doğrulamada veriseti eğitim ve test şeklinde ikiye ayrılmaz. Bunun yerine K değeri kaç seçildiyse o kadar eşit parçaya ayrılır ve K adımlık bir döngünün her adımında bu parçalardan biri test diğerleri (K-1

adeti) ise eğitim veriseti olarak kullanılır. Döngünün her adımında elde edilen skorlar ise kaydedilir ve döngü sonunda bu skorların ortalaması alınarak daha objektif bir başarı skoru elde edilmiş olunur. K-katlı çapraz doğrulama, veri madenciliği modellerini değerlendirmek için çok önemli bir tekniktir. Verimliliği artırır, varyansı azaltır ve önyargı ile varyans arasında dengeli bir denge sağlar (Mevik ve Cederkvist, 2004).

Tez çalışmasında kullanılan modeller ve seçilen parametreler sırasıyla aşağıda gösterilmiştir.

Şekil 2.16'da K-En yakın komşu algoritması temelli modelin parametre ekranı görülmektedir. Bu algoritma, Fix ve Hodges tarafından 1951 yılında geliştirilen parametrik olmayan bir denetimli öğrenme algortimasıdır (Fix ve Hodges, 1951). Hem sınıflandırma hem de regresyon problemlerinin çözümünde kullanılmaktadır. Bu algoritma regresyon problemlerinde kullanıldığında, en yakın k adet komşunun, mesafelerinin tersiyle ağırlıklandırılan bir ağırlıklı ortalama değerini kullanarak hedef veri noktasını tahmin etmeye çalışır.

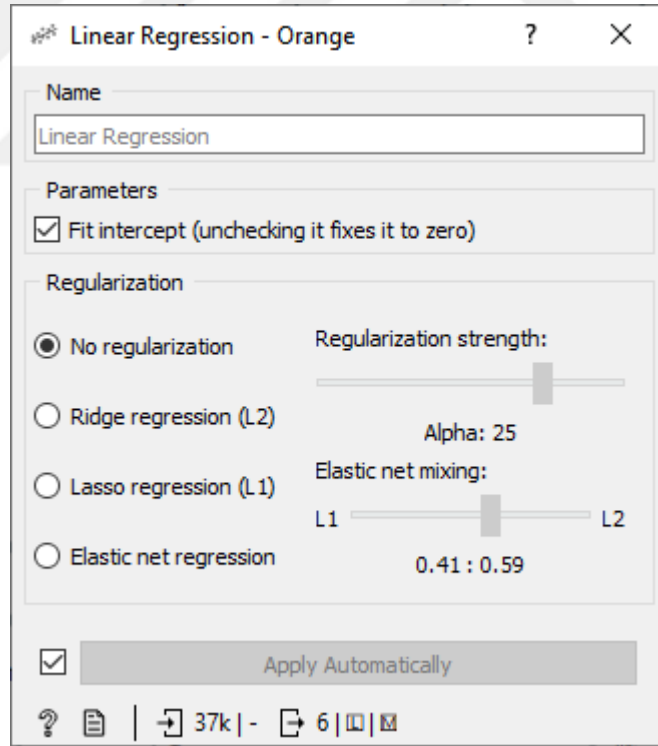


Şekil 2.16. K-En yakın komşu modeli parametreleri.

K-en yakın komşu modeli, öğrenme aşamasında eğitim verilerini depolar. Veri noktaları arasındaki uzaklığı belirlemek için bir mesafe metriği kullanılır (örneğin, Manhattan mesafesi). Tahmin yapılacak yeni bir veri noktası geldiğinde, bu noktanın diğer veri noktalarına olan uzaklıkları hesaplanır. En yakın k komşu belirlenir. Bu, tahmin yaparken kaç komşunun dikkate alınacağını gösterir. K-en yakın komşuların değerleri göz önüne alınarak, yeni veri noktasının tahmini değeri belirlenir. Bu genellikle komşu değerlerin ortalaması olabilir.

Algoritmanın kullanacağı mesafe ölçüm yöntemi “Metric” parametresi ile, ağırlıklandırma yöntemi ise “Weight” parametresi ile belirtilmektedir. “Number of neighbors” parametresi ise kaç adet yakın komşunun kullanılacağını belirtir.

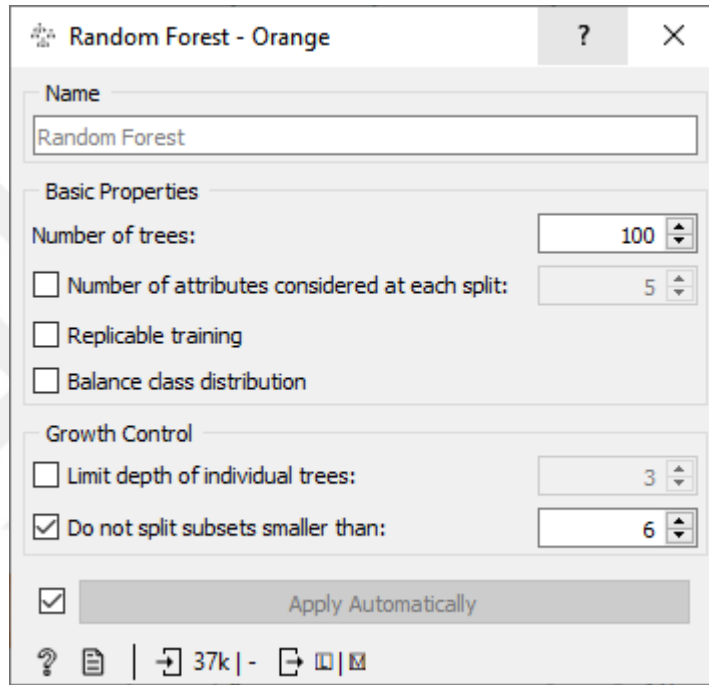
Şekil 2.17’de doğrusal regresyon algoritması temelli modelin parametre ekranı görülmektedir. Bu algoritma, verisetinde gözlemlenen gerçek değerler ile doğrusal yaklaşımla tahmin edilen değerler arasındaki ortalama karesel hatayı en aza indirmek için $w = (w_1, \dots, w_p)$ katsayılarına sahip doğrusal bir modele elde etmeye çalışır. Bu ekranda yer alan “Regularization” parametresi ile tahmin edilmeye çalışılan büyüklüğe etki eden faktörlerin hepsine eşit yaklaşmak (no-regularization), önemli olan faktöre daha fazla anlam yüklemek (lasso) ya da farklı önemde olan faktörler olsa da bunların arasında bir denge kurmak (ridge) gibi farklı yaklaşımlar güdülebilir. Bu ekrandaki “Elastic net regression” parametresi ise “lasso” ve “ridge” yöntemleri arasında bir denge gözetmek için tercih edilebilir.



Şekil 2.17. Doğrusal regresyon modeli parametreleri.

Şekil 2.18’de rastgele orman algoritması temelli modelin parametre ekranı görülmektedir. Hem sınıflandırma hem de regresyon problemlerinin çözümünde kullanılmaktadır. Rastgele orman modeli oluşturulurken, önceden belirlenen sayıda

karar ağacı her biri orijinal veri setinden oluşturulmuş alt kümeler üzerinde eğitilir. Kullanılacak değişkenler ise her birinde rastgele seçilir. Son tahminin oluşturulma şekli, rastgele orman modelinde çözülmeye çalışılan probleme göre değişir. Bu bir sınıflandırma problemi ise tüm karar ağaçlarından elde edilen tahminlerde oy çoğunluğuna bakılır; regresyon problemi ise tahmin olasılıklarının ortalaması alınır. Bu nedenle genellikle tek bir karar ağacına göre daha başarılı sonuçlar elde edilir. Şekil 2.18’de ki ekranda yer alan “Number of trees” parametresi rastgele orman içerisinde kullanılacak karar ağacı sayısını belirtir.

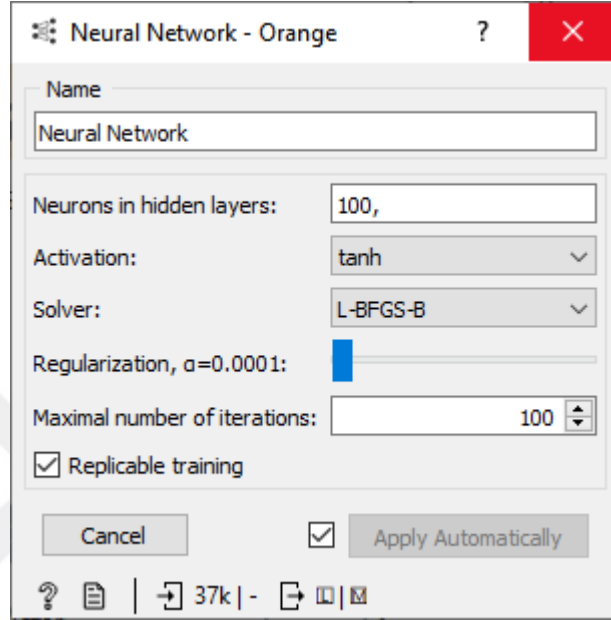


Şekil 2.18. Rastgele orman modeli parametreleri.

Şekil 2.19’da yapay sinir ağı algoritması temelli modelin parametre ekranı görülmektedir. Bu ekranda yapay sinir ağının içerisinde bulunan gizli katmanlarda yer alan nöron sayısı, aktivasyon fonksiyonu, çözüm yaklaşımı ve eğitimin en fazla kaç döngü adımıyla oluşacağı gibi ayarlamalar yapılabilmektedir. Yapay Sinir Ağları (YSA’lar), insan beyninden ilham alarak tasarlanmış, her biri kendi belleğine sahip, birbirine bağlı işlem birimlerinden oluşan paralel ve dağıtılmış bilgi işleme yapılarıdır (Elmas, 2011). Bu yapay zeka modelleri, biyolojik sinir hücrelerini (nöronları) taklit ederek oluşturulmuştur. Yapay sinir hücreleri arasındaki bağlantılar, biyolojik sinir ağlarındaki sinaptik bağlantılara benzer şekilde çalışarak bilgiyi işler ve depolar.

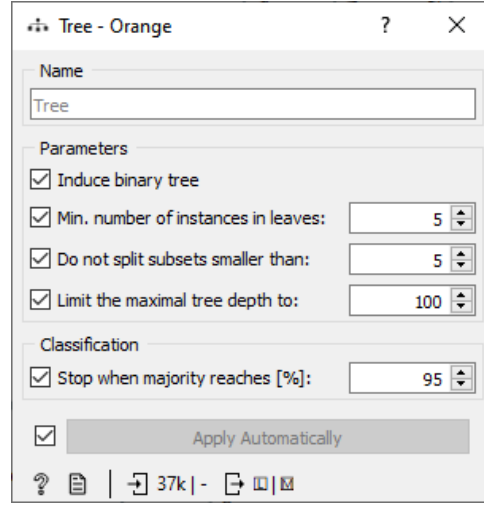
Yapay Sinir Ağları, doğadan esinlenerek geliştirilen yapay yapılardır ve insan beyninin öğrenme kapasitesini taklit edebilme özellikleri sayesinde geniş bir uygulama

alanına sahiptir. Bu modelleme yaklaşımı, özellikle desen tanıma, sınıflandırma ve öğrenme problemlerinde başarıyla kullanılarak karmaşık bilgi işleme görevlerini gerçekleştirebilir. Yapay Sinir Ağları, bilgi işleme konseptlerini biyolojik sistemlerden ödünç alarak, modern yapay zeka uygulamalarının temelini oluşturan önemli bir yapıdır.



Şekil 2.19. Yapay sinir ağı modeli parametreleri.

Şekil 2.20’de ise karar ağacı algoritması temelli modelin parametre ekranı görülmektedir. Makine öğreniminde karar ağacı, bir değişkene ait gözlemlerin hedef değişkende meydana getirdiği sonuçlarıyla eşleştirildiği bir modeldir. Karar ağaçları, yorumlanabilirliği, kullanım kolaylığı ve hem sayısal hem de kategorik verileri işleyebilme yeteneği nedeniyle yaygın olarak kullanılmaktadır. Hastalık tahmini, sahtekarlık tespiti ve öğrenci performansı tahmini gibi birçok alanda kullanılırlar ve sıklıkla rastgele ormanlar ve gradyan artırma gibi topluluk modellerinde kullanılırlar.

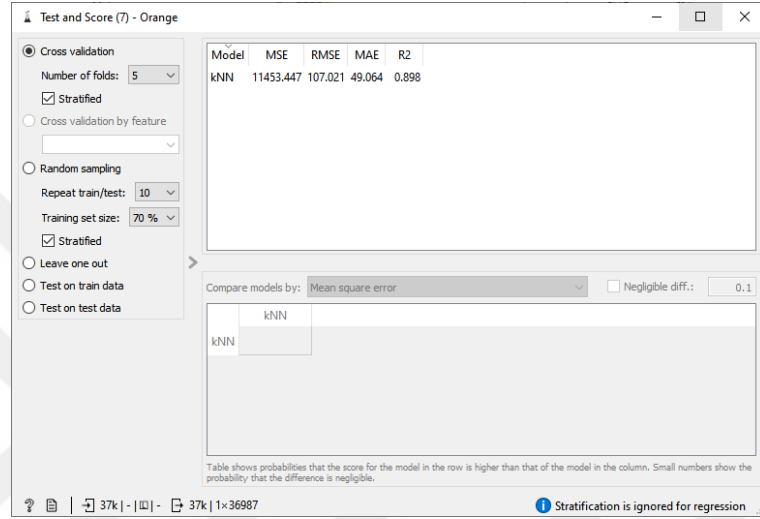


Şekil 2.20. Karar ağacı modeli parametreleri.

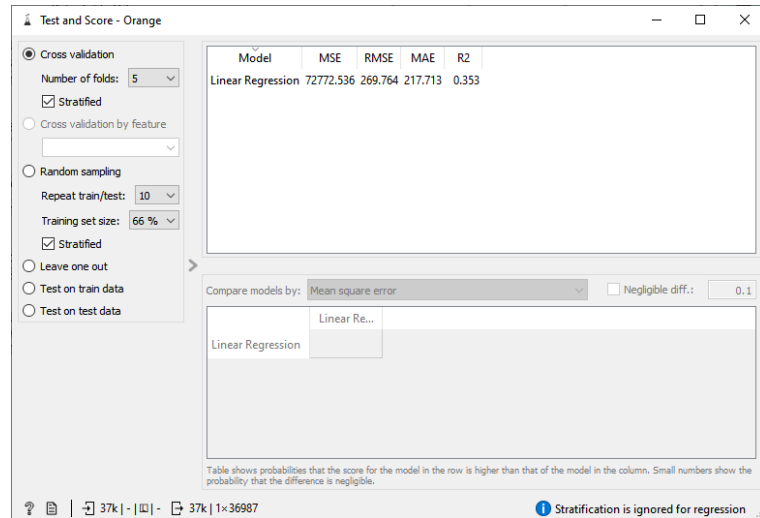
Kullanılan bu modellere ait sonuçlarla ilgili değerlendirmeler bir sonraki bölümde ele alınacaktır.

3. BULGULAR VE İRDELEME

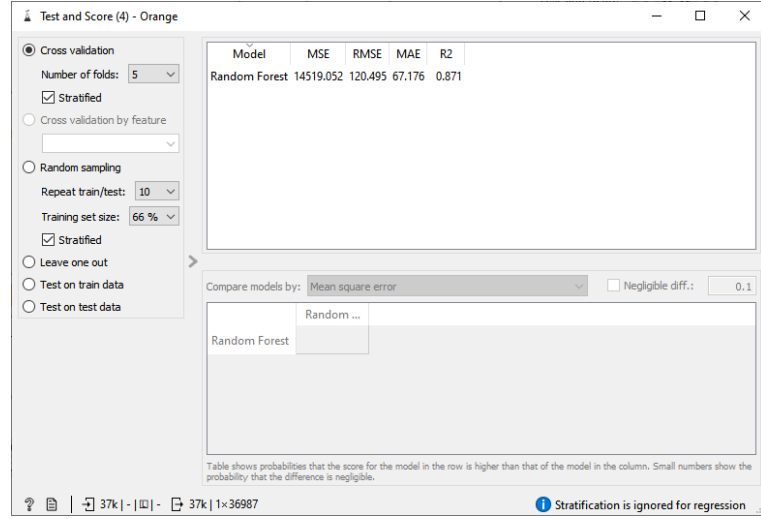
Orange veri madenciliği platformunda kullanılan modeller ile elde edilen sonuçlar aşağıda sırayla gösterilmiştir.



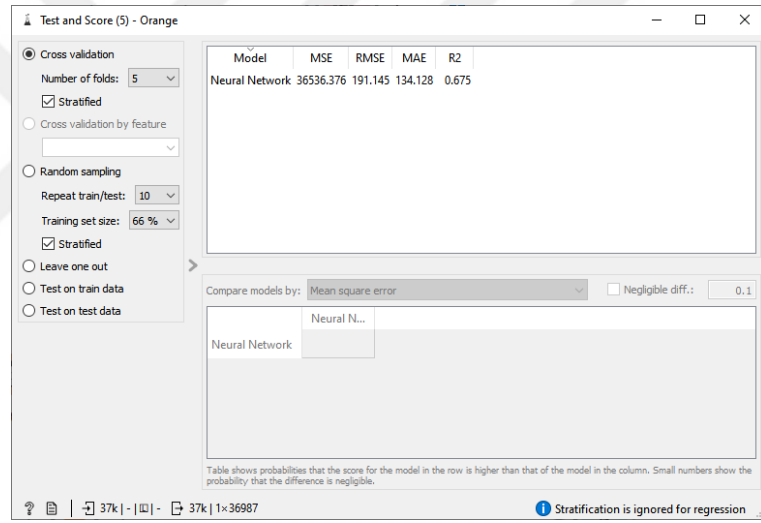
Şekil 3.1. K-En yakın komşu modeli sonuçları.



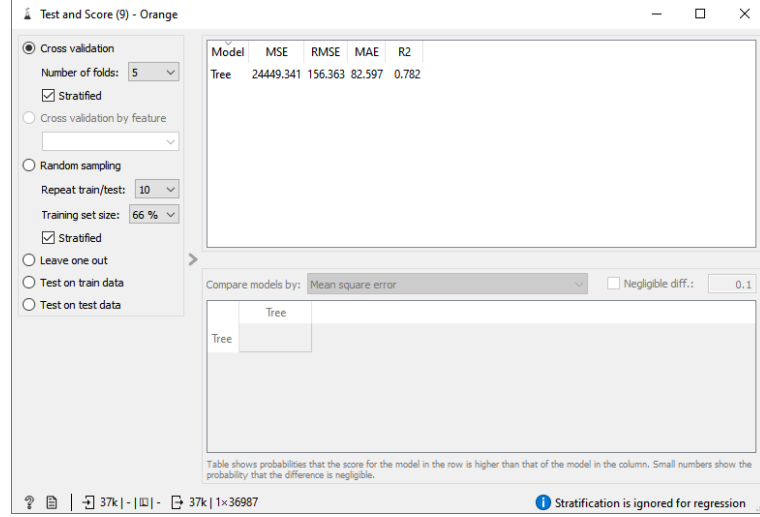
Şekil 3.2. Doğrusal regresyon modeli sonuçları.



Şekil 3.3. Rastgele orman modeli sonuçları.



Şekil 3.4. Yapay sinir ağı modeli sonuçları.



Şekil 3.5. Karar ağacı modeli sonuçları.

Sırasıyla Şekil 3.1, Şekil 3.2, Şekil 3.3, Şekil 3.4 ve Şekil 3.5’de gösterilen sonuçlar incelendiğinde aşağıdaki bulgular elde edilmiştir.

1. En düşük MSE, RMSE ve MAE skoruna, en yüksek R^2 skoruna sahip olan dolayısıyla en başarılı olan model K-En yakın komşu modeli olmuştur. Bu modelin MSE, RMSE, MAE ve R^2 skorları sırasıyla, 11.453,447, 107,021, 49,064, 0,898 olmuştur.
2. İkinci en başarılı model MSE, RMSE, MAE ve R^2 ölçütlerinde elde ettiği 14.519,052, 120,495, 67,176, 0,871 skorlarıyla Rastgele orman modeli olmuştur. Bu model içerisinde 100 adet karar ağacı barındırmaktadır.
3. Rastgele orman modeli içerisinde yer alan karar ağacı sayısı 100’den 200’e çıkarıldığında rastgele orman modeli MSE, RMSE, MAE ve R^2 ölçütlerinde elde ettiği 14.478,510, 120,327, 66,982, 0,871 skorlarını elde edebilmiştir. Sonuç olarak karar ağacı sayısının iki katına çıkartılması tahmin performansında beklenen artışı getirememiş, aksine modelin tahmin etme süresi çok fazla artmıştır.
4. Benzer şekilde rastgele orman modeli içerisinde yer alan karar ağacı sayısı 100’den 50’ye indirildiğinde ise rastgele orman modeli MSE, RMSE, MAE ve R^2 ölçütlerinde elde ettiği 14.743,283, 121,422, 67,613, 0,869 skorlarını elde edebilmiştir. Sonuç olarak karar ağacı sayısının yarıya indirilmesi tahmin performansında çok küçük bir azalma getirmiş olsa da modelin tahmin etme süresi çok daha kısalmıştır. Model eğitilme ve test edilme hızının daha önemli

olduğu senaryolarda daha az sayıda karar ağacı kullanmak tahmin performansında çok az bir düşme göze alınarak tercih edilebilir. Burada önemli olan hız ve doğruluk arasında bir denge yakalayabilmektir.

5. Karar ağacı modeli ise beklendiği gibi içerisinde çok daha fazla sayıda karar ağacı barındıran rastgele orman modelinden daha düşük performans göstermiştir. Karar ağacı modelinin MSE, RMSE, MAE ve R^2 ölçütlerinde elde ettiği skorlar sırasıyla 24.449,341, 156,363, 82,597, 0,782 olmuştur.
6. Yapay sinir ağı modeli ise karar ağacı modelinden daha düşük performans göstererek MSE, RMSE, MAE ve R^2 ölçütlerinde sırasıyla 36.536,376, 191,145, 134,128, 0,675 skorlarını elde edebilmiştir. Bu modelde gizli katmanlardaki nöron sayısı 100 olarak seçilmiştir.
7. Yapay sinir ağı modeli gizli katmanlarında yer alan nöron sayısı 100'den 200'e çıkarıldığında yapay sinir ağı orman modeli MSE, RMSE, MAE ve R^2 ölçütlerinde elde ettiği 34.989,430, 187,055, 129,344, 0,689 skorlarını elde edebilmiştir. Sonuç olarak nöron sayısının iki katına çıkartılması tahmin performansında çok az bir artış getirebilmiş fakat zaten diğer modellere göre oldukça uzun olan modelin tahmin etme süresi çok daha fazla artmıştır.
8. En başarısız model MSE, RMSE, MAE ve R^2 ölçütlerinde sırasıyla 72.772,536, 269,764, 217,713, 0,353 skorlarını elde edebilen Doğrusal regresyon modeli olmuştur. Bu modelin başarısız olmasının nedeni doğrusal regresyon algoritmasının doğrusal olmayan sistemleri gerektiği gibi modelleyememesidir. Tez çalışmasının konusu olan meteorolojik verilerdeki büyüklükler doğrusal olmayan bir sisteme aittirler. Literatüre göre meteorolojik sistemler karmaşık ve kaotik doğaları nedeniyle doğrusal olmayan sistemler olarak kabul edilir, bu da düzensizliğe ve daha az öngörülebilirliğe yol açar (Wang vd., 2016; Yan vd, 2021). Bu nedenle, meteorolojik sistemlerin doğrusal olmayan yapısı ve kaotik doğası, karmaşık davranışlarını ve bunun sonuçlarını yakalamak için ileri modelleme ve tahmin tekniklerinin kullanılmasını gerektirir (Pérez-Muñuzuri, 1998; Pérez-Muñuzuri ve Gelpi, 2000; Jin vd., 2019).

Orange veri madenciliği platformunda kullanılan modeller ile elde edilen sonuçlar aşağıdaki çizelgede tahmin başarı sırayla gösterilmiştir.

Çizelge 3.1. Orange veri madenciliği platformunda kullanılan modeller ile elde edilen sonuçlar.

Modeller	MSE	RMSE	MAE	R²
K-En Yakın Komşu Modeli	11.453,447	107,021	49,064	0,898
Rastgele Orman Modeli (K.A. 200)	14.478,510	120,327	66,982	0,871
Rastgele Orman Modeli (K.A. 100)	14.519,052	120,495	67,176	0,871
Rastgele Orman Modeli (K.A. 50)	14.743,283	121,422	66,982	0,869
Karar Ağacı Modeli	24.449,341	156,363	82,597	0,782
Yapay Sinir Ağları Modeli (N. 200)	34.989,430	187,055	129,344	0,689
Yapay Sinir Ağları Modeli (N. 100)	36.536,376	191,145	134,128	0,675
Doğrusal Regresyon Modeli	72.772,536	269,764	217,713	0,353

4. SONUÇLAR VE ÖNERİLER

Bu tez çalışmasında, bazı iklimsel veriler kullanılarak güneş ışıınım şiddetinin modellenmesi için Orange veri madenciliği platformunda yer alan bazı modeller kullanılmış, kullanılan modeller belirli yönleriyle optimize edilmeye çalışılmış ve elde edilen sonuçlar karşılaştırılmıştır. Çalışmada ele alınan problem tam anlamıyla bir regresyon problemi olup, çalışmada kullanılan veriler ise literatürde de belirtildiği gibi doğrusal olmayan bir sistemden gelmektedir. Orange veri madenciliği platformunda regresyon problemleri için kullanılabilen modellerden Yapay Sinir Ağları (YSA), Karar Ağacı (KA), Rastgele Orman (RO), Doğrusal Regresyon (DR) ve K-En Yakın Komşu (KEY) modelleri kullanılmıştır.

Çalışmada kullanılan veri seti, Oregon Üniversitesi yerleşkesinde NREL araştırmacıları tarafından kurulan ölçüm istasyonundan elde edilen iklimsel verileri içermektedir. Bu veriler, güneş ışıınım şiddeti ile ilişkilendirilen hava sıcaklığı, bağıl nem ve atmosferik basınç gibi değişkenleri içermektedir. Ayrıca, ham veri setinde elektriksel parametrelerin ölçümleri de bulunmaktadır. Ölçümler, 1 Ocak 2013 ile 31 Aralık 2013 tarihleri arasında gerçekleştirilmiştir. Verilerin analizi sırasında, aykırı değer içeren örnekler öncelikle temizlenmiş ve ardından iki yeni değişken türetilmiştir. Türetilen değişkenler, ölçüm yapılan zaman bilgisinden elde edilmiştir. Belirlenen modeller, 5-katlı çapraz doğrulama ile eğitilmiştir, ve modellerin performansları çeşitli ölçütler kullanılarak değerlendirilmiştir. Bu ölçütler arasında Ortalama Kare Hata (MSE), Kök Ortalama Kare Hata (RMSE), Ortalama Mutlak Hata (MAE) ve R2 skoru bulunmaktadır.

Bu süreçler, güneş ışıınım şiddetinin tahmininde kullanılan modellerin güvenilirliğini değerlendirmek amacıyla gerçekleştirilmiştir. Bu çalışma, iklimsel faktörlerin elektrik üretimi üzerindeki etkisini anlamak ve güneş enerjisi sistemlerinin performansını optimize etmek için önemli bir adımdır.

Tüm modellerden elde edilen sonuçlar karşılaştırıldığında regresyon problemine uygun olan ve olmayan modeller belirlenmiştir.

Ayrıca bu tez çalışmasıyla, Orange veri madenciliği platformunun herhangi bir programlama dili bilmeyi gerektirmeden, veri madenciliği, veri modelleme, makine

öğrenmesi ile ilgili çalışmalarda kullanılabileceği ve çok hızlı bir şekilde sonuçların elde edilebileceği gösterilmiştir. Benzer platformlar olan Weka® veri platformu ve Knime® analitik platformu ile de benzer sonuçlar alınabilmektedir.

Çalışmanın bundan sonrası için mimarisinde tıpkı rastgele orman modeli gibi çoklu karar ağacı kullanan gradyan arttırma (Gradient Boosting), kategorik arttırma (CatBoost), uyarlamalı arttırma (AdaBoost) gibi arttırma algoritması temelli modeller ile daha fazla gizli katmanlı YSA modelleri kullanılarak bu tez çalışmasında tercih edilen modellerle karşılaştırılabilir.



5. KAYNAKLAR

- Abid, M., Usman, M., & Ashraf, M. (2017). Plagiarism detection process using data mining techniques. *International Journal of Recent Contributions From Engineering Science & It (Ijes)*, 5(4), 68. <https://doi.org/10.3991/ijes.v5i4.7869>
- Almanza, V. and Batyrshin, I. (2011). On trend association analysis of time series of atmospheric pollutants and meteorological variables in mexico city metropolitan area., 95-102. https://doi.org/10.1007/978-3-642-21587-2_11
- Andrews, R., Wynn, M., Vallmuur, K., Hofstede, A., Bosley, E., Elcock, M., ... & Rashford, S. (2019). Leveraging data quality to better prepare for process mining: an approach illustrated through analysing road trauma pre-hospital retrieval and transport processes in queensland. *International Journal of Environmental Research and Public Health*, 16(7), 1138. <https://doi.org/10.3390/ijerph16071138>
- Arifin, B., Perwitasari, D., Zulkarnain, Z., & Rokhman, M. (2022). Models for predicting the quality of life domains on the general population through the orange data mining approach. *Pharmaciana*, 12(1), 72. <https://doi.org/10.12928/pharmaciana.v12i1.20827>
- Aulia, D. and Waspada, I. (2019). The design of exploratory application and preprocessing of event log data in lms moodle-based online learning activities for process mining. *Khazanah Informatika Jurnal Ilmu Komputer Dan Informatika*, 5(2), 124-133. <https://doi.org/10.23917/khif.v5i2.8023>
- Aziz, S. and Awlla, A. (2019). Performance analysis and prediction student performance to build effective student using data mining techniques. *Uhd Journal of Science and Technology*, 3(2), 10-15. <https://doi.org/10.21928/uhdjst.v3n2y2019.pp10-15>
- Babu, J., Niveditha, M., Bhavya, V., & Gowthami, K. (2017). Data mining techniques for herbs. *International Journal of Engineering & Technology*, 7(1.1), 406. <https://doi.org/10.14419/ijet.v7i1.1.9943>

- Baradwaj, B. and Pal, S. (2011). Mining educational data to analyze students performance. *International Journal of Advanced Computer Science and Applications*, 2(6). <https://doi.org/10.14569/ijacsa.2011.020609>
- Barjouei, A. and Naseri, M. (2021). A comparative study of statistical techniques for prediction of meteorological and oceanographic conditions: an application in sea spray icing. *Journal of Marine Science and Engineering*, 9(5), 539. <https://doi.org/10.3390/jmse9050539>
- Bartok, J., Habala, O., Bednar, P., Gazak, M., & Hluchý, L. (2010). Data mining and integration for predicting significant meteorological phenomena. *Procedia Computer Science*, 1(1), 37-46. <https://doi.org/10.1016/j.procs.2010.04.006>
- Beebe, N. and Clark, J. Dealing with terabyte data sets in digital investigations., 3-16. https://doi.org/10.1007/0-387-31163-7_1
- Bilalli, B., Abelló, A., & Aluja-Banet, T. (2017). On the predictive power of meta-features in openml. *International Journal of Applied Mathematics and Computer Science*, 27(4), 697-712. <https://doi.org/10.1515/amcs-2017-0048>
- Buizza, R., Houtekamer, P., Tóth, Z., Pellerin, G., Wei, M., & Zhu, Y. (2005). A comparison of the ecmwf, msc, and ncep global ensemble prediction systems. *Monthly Weather Review*, 133(5), 1076-1097. <https://doi.org/10.1175/mwr2905.1>
- Charles, A., Namen, A., & Rodrigues, P. (2017). Comparison of data mining models applied to a surface meteorological station. *RBRH*, 22(0). <https://doi.org/10.1590/2318-0331.0217170029>
- Chen, C., Xue, X., Sun, D., Zhao, R., Han, Y., Chen, T., ... & Zhao, Y. (2022). Comparison of lower stratosphere wind observations from the ustrc's rayleigh doppler lidar and the esa's satellite mission aeolus. *Earth and Space Science*, 9(8). <https://doi.org/10.1029/2021ea002176>
- Chen, Z. (2006). From data mining to behavior mining. *International Journal of Information Technology & Decision Making*, 05(04), 703-711. <https://doi.org/10.1142/s0219622006002271>
- Cuzzocrea, A. (2014). Models and algorithms for high-performance data management and mining on computational grids and clouds. *Journal of Grid Computing*, 12(3), 443-445. <https://doi.org/10.1007/s10723-014-9316-5>

- Demšar, J., Zupan, B., Leban, G., & Curk, T. (2004). Orange: from experimental machine learning to interactive data mining., 537-539. https://doi.org/10.1007/978-3-540-30116-5_58
- Eck, M., Lu, X., Leemans, S., & Aalst, W. (2015). Pm²: A process mining project methodology., 297-313. https://doi.org/10.1007/978-3-319-19069-3_19
- Eler, D., Grosa, D., Pola, I., Garcia, R., Correia, R., & Teixeira, J. (2018). Analysis of document pre-processing effects in text and opinion mining. *Information*, 9(4), 100. <https://doi.org/10.3390/info9040100>
- Elmas, Ç. (2011). *Yapay Zeka Uygulamaları*. Seçkin Yayıncılık.
- Fix, E., Hodges, J. L. (1951). Discriminatory Analysis. Nonparametric Discrimination: Consistency Properties (PDF) (Report). USAF School of Aviation Medicine, Randolph Field, Texas. Archived (PDF) from the original on September 26, 2020.
- Fletcher, S. and Islam, Z. (2014). Measuring information quality for privacy preserving data mining. *International Journal of Computer Theory and Engineering*, 7(1), 21-28. <https://doi.org/10.7763/ijcte.2015.v7.924>
- Folorunsho, O. and Adeyemo, A. (2012). Application of data mining techniques in weather prediction and climate change studies. *International Journal of Information Engineering and Electronic Business*, 4(1), 51-59. <https://doi.org/10.5815/ijieeb.2012.01.07>
- Gandhe, S., Talele, K., & Keskar, (2007). Image mining using wavelet transform., 797-803. https://doi.org/10.1007/978-3-540-74819-9_98
- Godec, P., Pančur, M., Ilenič, N., Čopar, A., Stražar, M., Erjavec, A., ... & Zupan, B. (2019). Democratized image analytics by visual programming through integration of deep models and small-scale machine learning. *Nature Communications*, 10(1). <https://doi.org/10.1038/s41467-019-12397-x>
- Guo, X., Sun, X., & Ma, J. (2011). Prediction of daily crop reference evapotranspiration (et₀) values through a least-squares support vector machine model. *Hydrology Research*, 42(4), 268-274. <https://doi.org/10.2166/nh.2011.072>

- Habala, O., Hluchy, L., Tran, V., Krammer, P., & Seleng, M. (2012). Using advanced data mining and integration in environmental prediction scenarios. *Computer Science*, 13(1), 5. <https://doi.org/10.7494/csci.2012.13.1.5>
- Hasan, R., Palaniappan, S., Mahmood, S., Abbas, A., Sarker, K., & Sattar, M. (2020). Predicting student performance in higher educational institutions using video learning analytics and data mining techniques. *Applied Sciences*, 10(11), 3894. <https://doi.org/10.3390/app10113894>
- He, W. (2013). Examining students' online interaction in a live video streaming environment using data mining and text mining. *Computers in Human Behavior*, 29(1), 90-102. <https://doi.org/10.1016/j.chb.2012.07.020>
- Houghton, I., Hegermiller, C., Teicheira, C., & Smit, P. (2022). Operational assimilation of spectral wave data from the sofar spotter network. *Geophysical Research Letters*, 49(15). <https://doi.org/10.1029/2022gl098973>
- Huertas, J., Huertas, M., & Díaz, J. (2012). Assessing precision and accuracy of atmospheric emission inventories. *International Journal of Environmental Science and Technology*, 9(2), 195-202. <https://doi.org/10.1007/s13762-012-0022-1>
- Ismail, M., Ibrahim, M., Sanusi, Z., & Nat, M. (2015). Data mining in electronic commerce: benefits and challenges. *International Journal of Communications Network and System Sciences*, 08(12), 501-509. <https://doi.org/10.4236/ijcns.2015.812045>
- Jabbar, M., Deekshatulu, B., & Chandra, P. (2013). Classification of heart disease using k- nearest neighbor and genetic algorithm. *Procedia Technology*, 10, 85-94. <https://doi.org/10.1016/j.protcy.2013.12.340>
- Jiménez, P. and Dudhia, J. (2012). Improving the representation of resolved and unresolved topographic effects on surface wind in the wrf model. *Journal of Applied Meteorology and Climatology*, 51(2), 300-316. <https://doi.org/10.1175/jamc-d-11-084.1>
- Jin, Q., Fan, X., Liu, J., Xue, Z., & Jian, H. (2019). Using extreme gradient boosting to predict changes in tropical cyclone intensity over the western north pacific. *Atmosphere*, 10(6), 341. <https://doi.org/10.3390/atmos10060341>

Karadimce, A. (2015). Model of cloud-based services for data mining analysis. *Computer and Information Science*, 8(4), 40. <https://doi.org/10.5539/cis.v8n4p40>

Kaul, S. (2017). Review of existing data mining techniques used for weather prediction. *International Journal for Research in Applied Science and Engineering Technology*, V(XI), 2530-2537. <https://doi.org/10.22214/ijraset.2017.11353>

Khan, I., Zhu, H., Yao, J., Khan, D., & Iqbal, T. (2017). Hybrid power forecasting model for photovoltaic plants based on neural network with air quality index. *International Journal of Photoenergy*, 2017, 1-9. <https://doi.org/10.1155/2017/6938713>

Kia, S., Nambiar, M., Thé, J., Gharabaghi, B., & Aliabadi, A. (2022). Machine learning to predict area fugitive emission fluxes of ghgs from open-pit mines. *Atmosphere*, 13(2), 210. <https://doi.org/10.3390/atmos13020210>

Kumar, A. (2017). A study on web content mining. *International Journal of Engineering and Computer Science*. <https://doi.org/10.18535/ijecs/v6i1.29>

Liu, H. and Wang, B. (2007). An association rule mining algorithm based on a boolean matrix. *Data Science Journal*, 6, S559-S565. <https://doi.org/10.2481/dsj.6.s559>

Liu, N., Gao, G., & Liu, G. (2016). Data preprocessing based on partially supervised learning.. <https://doi.org/10.2991/icimm-16.2016.121>

Lu, Q. (2011). Initial evaluation and assimilation of fy-3a atmospheric sounding data in the ecmwf system. *Science China Earth Sciences*, 54(10), 1453-1457. <https://doi.org/10.1007/s11430-011-4243-9>

Ma, J., Zhang, R., Li, L., Liu, Z., Sun, J., Kong, L., ... & Liu, X. (2022). Analysis of the dust-concentration distribution law in an open-pit mine and its influencing factors. *Acs Omega*, 7(48), 43609-43620. <https://doi.org/10.1021/acsomega.2c04439>

Mahatme, V. and Bhoyar, K. (2018). Impact of distance metrics on the performance of k-means and fuzzy c-means clustering – an approach to assess student's performance in e-learning environment. *International Journal of Advanced Research in Computer Science*, 9(1), 887-892. <https://doi.org/10.26483/ijarcs.v9i1.5417>

Maimon, O. and Rokach, L. (2009). Introduction to knowledge discovery and data mining., 1-15. https://doi.org/10.1007/978-0-387-09823-4_1

- Marion, B., Anderberg, A., Deline, C., del Cueto, J., Muller, M., Perrin, G., ... King, B. (2014). New data set for validating PV module performance models. In 2014 IEEE 40th Photovoltaic Specialist Conference (PVSC) (pp. 1362–1366). IEEE.
<https://doi.org/10.1109/PVSC.2014.6925171>
- Mariscal, G., Marbán, Ó., & Fernandez, C. (2010). A survey of data mining and knowledge discovery process models and methodologies. *The Knowledge Engineering Review*, 25(2), 137-166. <https://doi.org/10.1017/s0269888910000032>
- Menzies, T. (2013). Beyond data mining. *Ieee Software*, 30(3), 92-92.
<https://doi.org/10.1109/ms.2013.49>
- Mevik, B. and Cederkvist, H. (2004). Mean squared error of prediction (mse) estimates for principal component regression (pcr) and partial least squares regression (pls). *Journal of Chemometrics*, 18(9), 422-429.
<https://doi.org/10.1002/cem.887>
- Mr, M. (2021). Analysis on data mining tools used in business intelligence and inventions. *Journal of Computing and Natural Science*, 57-68.
<https://doi.org/10.53759/181x/jcns202101010>
- Mucherino, A., Papajorgji, P., & Pardalos, P. (2009). A survey of data mining techniques applied to agriculture. *Operational Research*, 9(2), 121-140.
<https://doi.org/10.1007/s12351-009-0054-6>
- Nugroho, S. and Prihandoko, P. (2018). Architecture of text mining application in analyzing public sentiments of west java governor election using naive bayes classification. *International Journal of Computer Applications*, 182(13), 1-5.
<https://doi.org/10.5120/ijca2018916916>
- Obelets, T. (2022). Data mining tools for complex socio-economic processes and systems. *System Research and Information Technologies*, (4), 68-78.
<https://doi.org/10.20535/srit.2308-8893.2022.4.06>
- Olaleye, D. and Tardiff, B. (2001). Practical issues in and applications of clinical data mining. *Drug Information Journal*, 35(3), 791-808.
<https://doi.org/10.1177/009286150103500317>

- Patel, K. and Patel, A. (2012). Process of web usage mining to find interesting patterns from web usage data. *International Journal of Computers & Technology*, 3(1), 144-148. <https://doi.org/10.24297/ijct.v3i1c.2767>
- Patil, P., Thube, S., Ratnaparkhi, B., & Rajeswari, K. (2014). Analysis of different data mining tools using classification, clustering and association rule mining. *International Journal of Computer Applications*, 93(8), 35-39. <https://doi.org/10.5120/16238-5766>
- Pérez-Muñuzuri, V. (1998). Forecasting of chaotic cloud absorption time series for meteorological and plume dispersion modeling. *Journal of Applied Meteorology*, 37(11), 1430-1443. [https://doi.org/10.1175/1520-0450\(1998\)0372.0.co;2](https://doi.org/10.1175/1520-0450(1998)0372.0.co;2)
- Pérez-Muñuzuri, V. and Gelpi, I. (2000). Application of nonlinear forecasting techniques for meteorological modeling. *Annales Geophysicae*, 18(10), 1349-1359. <https://doi.org/10.1007/s005850000263>
- Pérez-Sánchez, D. and Thorne, M. (2014). Investigation into the transport of 238u-series radionuclides in soils to plants., 665-672. https://doi.org/10.1007/978-3-319-11059-2_76
- Pessoa, A., Lima, G., Silva, J., Stephany, S., Strauss, C., Caetano, M., ... & Ferreira, N. (2012). Mineração de dados meteorológicos para previsão de eventos severos. *Revista Brasileira De Meteorologia*, 27(1), 61-74. <https://doi.org/10.1590/s0102-77862012000100007>
- Podpečan, V., Zemenova, M., & Lavrač, N. (2011). Orange4ws environment for service-oriented data mining. *The Computer Journal*, 55(1), 82-98. <https://doi.org/10.1093/comjnl/bxr077>
- Rahim, S., Rahman, M., & Chowdhury, A. (2017). Mining industrial engineered data of apparel industry: a proposed methodology. *International Journal of Computer Applications*, 161(7), 1-7. <https://doi.org/10.5120/ijca2017913262>
- Sagayaraj, S. and N, V. (2018). Improving air quality management using gradient boosting based hierarchical temporal memory neural networks and fuzzy based classification based regression tree. *International Journal of Engineering & Technology*, 7(4), 12. <https://doi.org/10.14419/ijet.v7i2.9.9229>

- Saifuddin, S. and Ujianto, E. (2022). Classification of research proposal funding using naïve bayes and decision tree methods. *Journal Research of Social Science Economics and Management*, 2(03). <https://doi.org/10.36418/jrssem.v2i03.273>
- Sajwan, V., Awasthi, M., Goel, A., & Sharma, P. (2023). Sentiment analysis of twitter data regarding the agnipath scheme of the defense forces. *Indonesian Journal of Electrical Engineering and Computer Science*, 30(3), 1643. <https://doi.org/10.11591/ijeecs.v30.i3.pp1643-1650>
- Sasikala, M., Deepika, M., & Shankar, M. (2018). Pattern identification and predictions in data analysis. *International Journal of Engineering and Computer Science*, 7(03), 23686-23691. <https://doi.org/10.18535/ijecs/v7i3.05>
- Shamy, S., Luo, J., & Farag, T. (2015). Interactive framework for feasibility studies using data mining. *International Journal of Hybrid Information Technology*, 8(6), 115-124. <https://doi.org/10.14257/ijhit.2015.8.6.12>
- Sharma, G. (2020). Tendency of educational data mining in digital learning platform. *Current Trends in Computer Sciences & Applications*, 2(1). <https://doi.org/10.32474/ctcsa.2020.02.000127>
- Shelke, S. and Bhagat, B. (2015). Techniques for privacy preservation in data mining. *International Journal of Engineering Research And*, V4(10). <https://doi.org/10.17577/ijertv4is100473>
- Spruit, M., Kais, M., & Menger, V. (2021). Automated business goal extraction from e-mail repositories to bootstrap business understanding. *Future Internet*, 13(10), 243. <https://doi.org/10.3390/fi13100243>
- Stocker, J., Carruthers, D., Johnson, K., Hunt, J., & Chan, P. (2016). Optimized use of real-time vertical-profile wind data and fast modelling for prediction of airflow over complex terrain. *Meteorological Applications*, 23(2), 182-190. <https://doi.org/10.1002/met.1544>
- Stocker, J., Johnson, K., Forsyth, E., Smith, S., Gray, S., Carruthers, D., ... & Chan, P. (2022). Derivation of high-resolution meteorological parameters for use in airport wind shear now-casting applications. *Atmosphere*, 13(2), 328. <https://doi.org/10.3390/atmos13020328>

- Stovern, M., Guzman, H., Rine, K., Felix, O., King, M., Ela, W., ... & Sáez, A. (2016). Windblown dust deposition forecasting and spread of contamination around mine tailings. *Atmosphere*, 7(2), 16. <https://doi.org/10.3390/atmos7020016>
- Suguna, K. and Nandhini, K. (2017). Frequent pattern mining of web log files working principles. *International Journal of Computer Applications*, 157(3), 1-5. <https://doi.org/10.5120/ijca2017912642>
- Tokoro, C., Fukaki, K., Kadokura, M., & Fuchida, S. (2020). Forecast of amd quantity by a series tank model in three stages: case studies in two closed japanese mines. *Minerals*, 10(5), 430. <https://doi.org/10.3390/min10050430>
- Ueyama, H. (2023). Radiative cooling scale method for correcting hourly surface air temperature error in numerical weather prediction models. *Meteorological Applications*, 30(1). <https://doi.org/10.1002/met.2113>
- Ueyama, H., Kato, M., Kawamukai, H., Ikegami, M., & Fujimoto, H. (2015). Development of daily mean air temperature data with 50-m resolution for an information system identifying the suitable planting period for yamadanishiki sake rice. *Journal of Agricultural Meteorology*, 71(4), 292-301. <https://doi.org/10.2480/agrmet.d-14-00043>
- Ueyama, H., Shibata, S., Okada, S., & Ooki, H. (2018). Farmland meteorology estimation system for precise cultivation management. *Climate in Biosphere*, 18(0), 76-85. <https://doi.org/10.2480/cib.j-18-042>
- Vidyadevi, R. (2015). Secure protocol for mining in horizontally scattered database using association rule. *International Journal on Recent and Innovation Trends in Computing and Communication*, 3(2), 676-679. <https://doi.org/10.17762/ijritcc2321-8169.150252>
- Wang, H. and Wang, S. (2008). A knowledge management approach to data mining process for business intelligence. *Industrial Management & Data Systems*, 108(5), 622-634. <https://doi.org/10.1108/02635570810876750>
- Wang, J., Hu, X., Hollister, K., & Zhu, D. (2009) A comparison and scenario analysis of leading data mining software.. <https://doi.org/10.4018/9781605660608.ch032>

- Wang, J., Zhou, X., Shan-xin, Y., Li, B., & Yan, L. (2021). Analysis and correction of meteorological disturbance observed by ground radars in complex environment. *Plos One*, 16(10), e0258168. <https://doi.org/10.1371/journal.pone.0258168>
- Wang, Y., Nakakita, E., Kim, S., & Yamaguchi, K. (2016). Improving the accuracy of flood forecasting with transpositions of ensemble nwp rainfall fields considering orographic effects. *Journal of Hydrology*, 539, 345-357. <https://doi.org/10.1016/j.jhydrol.2016.05.047>
- Wasan, S., Bhatnagar, V., & Kaur, H. (2006). The impact of data mining techniques on medical diagnostics. *Data Science Journal*, 5, 119-126. <https://doi.org/10.2481/dsj.5.119>
- Wegener, D. and Rüping, S. (2010). On integrating data mining into business processes., 183-194. https://doi.org/10.1007/978-3-642-12814-1_16
- Wegener, D. and Rüping, S. (2011). Integration and reuse of data mining in business processes â a pattern-based approach. *International Journal of Business Process Integration and Management*, 5(3), 218. <https://doi.org/10.1504/ijbpim.2011.042526>
- Yan, B., Chan, P., Li, Q., He, Y., & Shu, Z. (2021). Dynamic analysis of meteorological time series in hong kong: a nonlinear perspective. *International Journal of Climatology*, 41(10), 4920-4932. <https://doi.org/10.1002/joc.7106>
- Yang, H. and Gao, Y. (2010). Research on similarity mining for flight data.. <https://doi.org/10.1109/etcs.2010.340>
- Yang, Y., Chen, R., Han, C., Liu, Z., & Wang, X. (2021). Optimal selection of empirical reference evapotranspiration method in 36 different agricultural zones of china. *Agronomy*, 12(1), 31. <https://doi.org/10.3390/agronomy12010031>
- Yu, X., Shi, Y., Zhang, L., Nie, G., & Huang, A. (2014). Intelligent knowledge beyond data mining: influences of habitual domains. *Communications of the Association for Information Systems*, 34. <https://doi.org/10.17705/1cais.03453>
- Zanev, V., Topalov, S., & Christov, V. (2014). Analysis and data mining of lead-zinc ore data. *Serdica Journal of Computing*, 7(3), 271-280. <https://doi.org/10.55630/sjc.2013.7.271-280>

Zeng, L., Chen, Q., & Huang, M. (2022). Rsfd: a rough set-based feature discretization method for meteorological data. *Frontiers in Environmental Science*, 10. <https://doi.org/10.3389/fenvs.2022.1013811>

Zulham, Z., Putri, E., & Hasugian, B. (2022). Pattern analysis of drug procurement system with fp-growth algorithm. *Jurnal Online Informatika*, 7(1), 70. <https://doi.org/10.15575/join.v7i1.841>



6. ÖZGEÇMİŞ

Kişisel Bilgiler

Adı Soyadı : S****k K**T
Uyruğu : T.C.
Doğum Yeri, Tarihi : E*****m, 0*/**/19**
Medeni Hali : ****
E-posta : s*****@mu.edu.tr

Eğitim

Alınan Derece	Aldığı Kurum/Üniversite	Mezuniyet Yılı
Lise	Ereğli Lisesi	1995
Lisans	Muğla Üniversitesi İstatistik ve Bilgisayar Bilimleri	2002

İş Tecrübesi

Yıl	Yer	Pozisyon/görev
2005 - 2018	Muğla Sıtkı Koçman Üniversitesi Bilgi İşlem Dairesi	Öğretim Görevlisi
2018 - 2020	Muğla Sıtkı Koçman Üniversitesi Uzaktan Eğitim Uygulama ve Araştırma Merkezi	Öğretim Görevlisi
2020 - 2024	Muğla Sıtkı Koçman Üniversitesi Bilgi İşlem Dairesi	Öğretim Görevlisi