Journal of Investigations on Engineering&Technology
Volume 2, Issue1, 31.08.2019
ISSN : 2687-3052

Yapay Zekâ Tabanlı Büyük Veri Yönetim Aracı

Fatih ARSLAN*[1], Hamdi Tolga KAHRAMAN[1a]

[1] Karadeniz Teknik Üniversitesi, Teknoloji Fakültesi, Yazılım Mühendisliği Bölümü, 61830, Trabzon

(Alınış / Received: 06.08.2019, Kabul / Accepted: 22.08.2019, Online Yayınlanma / Published Online: 31.08.2019)

Anahtar Kelimeler
Yapay Zekâ
Eksik Veri Tamamlama
Gürültü Tespiti ve Onarımı
Yapay Sinir Ağları
k-En Yakın Komşular(KNN)
Sezgisel KNN

Özet: Yapay zekâ, günümüzde birçok problemin çözüme kavuşturulmasında başrol oynamaktadır. Şu an ki konum itibariyle en değerli maden haline gelen veri, bilginin oluşmasındaki asıl kaynaktır. Bilgiyi elde etme süreci göz önüne alındığında, kaliteli bir bilgiyi elde etmek için ise verinin incelenmesi, analiz edilmesi ve işlenmeye hazır hale getirilmesi gerekmektedir. Veri analiz sürecinde karşılaşılan problemlerin başında, eksik ya da gürültülü/hatalı verilerin tespit edilmesi ve düzeltilmesi gelmektedir. Bu çalışmada eksik ve gürültülü verilerin tespit edilmesi ve düzeltilmesi amacıyla yapay zekâ tabanlı çalışan, özgün ve güçlü bir veri yönetim aracı geliştirilmiştir. Bu araç sayesinde veri setlerinin analiz edilmesi, bu veri setlerindeki eksik ve gürültülü verilerin tespit edilmesi ve düzeltilmesi sağlanacaktır. Geliştirilecek yazılım aracının özgünlüğü ise eksik ve gürültülü verileri düzeltme sürecinde modern, güçlü ve melez yapay zekâ algoritmalarını kullanacak olmasıdır.

Artificial Intelligence Based Big Data Management Tool

Keywords
Artificial intelligence
Missing Data Completion
Noise Detection and Repair
Artificial neural networks
K nearest neighborhood(KNN)
Intuitive KNN

Abstract: Today, artificial intelligence plays a leading role in solving many problems. Data, which has become the most valuable mine in terms of the current location, is the main source of information. When the process of obtaining information is taken into consideration, in order to obtain a quality information, the data must be examined, analyzed and made ready for processing. One of the problems encountered during the data analysis process is the identification and correction of missing or noisy / incorrect data. In this study, a unique and powerful data management tool based on artificial intelligence will be developed in order to detect and correct missing and noisy data. With this tool, data sets will be analyzed, missing and noisy data in these data sets will be detected and corrected. The originality of the software tool to be developed is that it will use modern, powerful and hybrid artificial intelligence algorithms in the process of correcting missing and noisy data.

*İlgili yazar: fatih.46ars@gmail.com ORCID: 0000-0001-6792-9293
[a] ORCID: 0000-0001-9985-6324

1. Giriş

Teknolojik gelişmelerin en hızlı yaşandığı ve hissedildiği alanların başında yapay zekâ gelmektedir. Yapay zekânın gelişimi üzerinde etkili olan başlıca öğelerden biri ise veridir. Veri, geçmişten günümüze bilginin oluşmasındaki asıl kaynaktır. İnsanın bir hücresinde tutulan verinin terabaytlar mertebesinde olduğu düşünüldüğünde biyolojik olarak insanın veri depolama kapasitesinin günümüz teknolojisinden üstün olduğu söylenebilir. Bunun yanında veri işleme teknolojisinde son yıllarda yaşanan gelişmeleri göz ardı etmek mümkün değildir. Çok çekirdekli, paralel işlem yapabilen, yüksek hızlı ve büyük kapasiteli yeni nesil işlemciler yanında veriyi bilgiye dönüştürmede insan ve diğer canlılara benzer karar mekanizmaları kullanan algoritmalar sayesinde çok güçlü veri madenciliği araçları geliştirilmektedir. Şirketler ürünlerinden daha fazla kazanç elde edebilmek, çalışmalarında verimliliği artırmak ve karar mekanizmalarını güçlendirebilmek için veri madenciliği için geliştirilmiş araçlardan faydalanmaktadırlar. Bu süreçte veri kalitesi ve bütünlüğü için harcadıkları zamanı ve bütçeyi artırmaktadırlar. Bunun nedeni, veri kalitesi ve bütünlüğünün firmalara katmakta olduğu değerin fark yaratmasıdır [1].

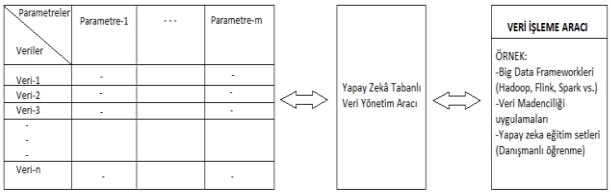
Yapay zekânın bir alt disiplini olan veri madenciliği teknolojik gelişmelerden etkilenerek büyük veri madenciliğine evrilmiştir. Bu değişimin temelinde ise internet-tabanlı alış-verişten uzay araçları ile yapılan haberleşmeye, endüstriyel otomasyon uygulamalarından sosyal medya uygulamalarına kadar sayısız alanda büyük bir veri yığının ortaya çıkması ve bunların elektronik ortamda saklanması ihtiyacı gelmektedir. Yarı iletken teknolojisindeki gelişmeler, veriyi depolama ve işleme kapasitesi yüksek elektronik malzemelerin üretilmesini sağlamıştır. Verideki ve işlem yapma kapasitesindeki artış ise veriyi işleyerek anlamlı bilgiye dönüştürmede kullanılan yapay zekâ algoritmalarının tekrar gözden geçirilmesine neden olmuştur. Geçmişte, günümüze kıyasla küçük sayılabilecek veri yığınları üzerinde etkili sonuçlar üreten algoritmalar bugün ise büyük veri işleme gereksinimini aynı şekilde karşılayamamaktadırlar. Günümüzde büyük veriyi işleyerek en kısa sürede en uygun sonucu bulmak giderek zor bir hale gelmiştir. En uygun sonuç, kabul edilebilir bir maliyet ile en kaliteli olan sonuçtur. Başarılı bir veri madenciliği için veriyi işleyecek algoritmadan daha önemlisi verinin kendisidir. Verinin, problem uzayını homojen bir şekilde örneklemesi ve doğru olması gerekir. Yani başarılı bir veri madenciliği etkisi elde etmek için kaliteli bir veriye ihtiyaç vardır. Kaliteli bir veriyi elde etmek için ise verinin incelenmesi, analiz edilmesi ve işlenmeye hazır hale getirilmesi gerekir. Veri analiz süreci, bilimsel araştırma sürecinin ve veri madenciliğinin en önemli basamaklarından biridir. Bu süreçte toplanan veriler amaca en uygun istatiksel, matematiksel ya da yapay zekâ teknikleriyle işlenir veya analiz edilir.

Veri analiz sürecinde karşılaşılan problemlerin başında, eksik ya da gürültülü/hatalı verilerin tespit edilmesi ve düzeltilmesi gelmektedir. Bu sorunun çözümünde kullanılmak üzere geçmişten günümüze farklı yöntemler geliştirilmiştir. Eksik veri ile analize devam etme, eksik gözlemleri analiz dışı bırakma, eksik gözlemler yerine veri atama veya çeşitli istatistiksel yöntemlerle eksik verileri tamamlama gibi yöntemler bu durumlarda sıkça kullanılmaktadırlar [2]. Bu yöntemler içerisinde araştırmacılar tarafından en çok kullanılan yöntemler, liste bazında silme ve çiftler bazında silme gibi eksik verileri analiz dışı bırakma yöntemleridir. Ancak yapılan çalışmalar bu yöntemlerin örneklemde kayba, güvenirlikte azalmaya, tahminlerde yanlılığa neden olduğunu [3] ve yanlılıktan kaynaklı olarak da örneklemin evreni temsil etme derecesinin düştüğünü göstermektedir [4]. Belirtilen bu sebeplerden dolayı, son yıllarda, bu yöntemler yerine, beklenti maksimizasyonu ve çoklu atama gibi modern yöntemler önerilmektedir. Çünkü bu yöntemler, silme yöntemleri gibi geleneksel kayıp veri yöntemlerinin aksine, yanlılığın azaltılması, etkili parametre tahminlerinin yapılması ve daha büyük istatistiksel gücün sağlanması hususunda daha etkili sonuçlar vermektedir [2].

Bu proje çalışmasında eksik ve gürültülü verilerin tespit edilmesi ve düzeltilmesi amacıyla yapay zekâ tabanlı çalışan, özgün ve güçlü bir veri yönetim aracı geliştirilmiştir. Bu araç sayesinde veri setlerinin analiz edilmesi, bu veri setlerindeki eksik ve gürültülü verilerin tespit edilmesi ve düzeltilmesi sağlanmaktadır. Geliştirilmiş olan yazılım aracının özgünlüğü ise eksik ve gürültülü verileri düzeltme sürecinde modern, güçlü ve melez yapay zekâ algoritmalarını kullanmış olmasıdır. Bu süreçte sezgisel optimizasyon algoritmalarından (genetik algoritma [5], yapay arı kolonisi algoritması [6], ortak yaşayan organizmalar [7] ve meta-sezgisel tahmin ve sınıflandırma algoritmalarından [8-9]) faydalanılmaktadır. Geliştirilmiş uygulamanın basit bir ara yüz ile kullanılması ve problemlere ait veri setlerinin kolaylıkla düzenlenebilmesi sağlanmaktadır. Veri setindeki eksiklikler ve gürültülü veriler veri yönetim yazılımı tarafından otomatik olarak tespit edilip araştırmacılara rapor halinde sunulmaktadır. Hata tespiti yapıldıktan sonra, program hatalı verilerin yerine geçebilecek en uygun değerleri bulup değiştirme işlemi yapmaktadır. Bu süreçte melez bir tahmin ve sınıflandırma tekniği olan "meta-sezgisel k-nn algoritması" ve doğrusal olmayan regresyon problemlerinin çözümlenmesi amacıyla literatürde en yaygın kullanılan algoritma olan yapay sinir ağları kullanılmıştır.

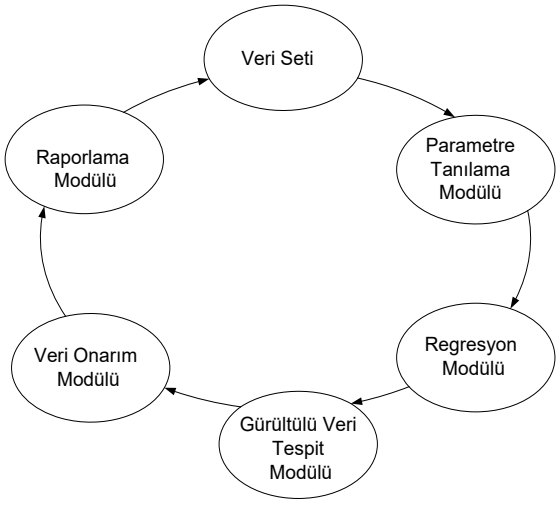
2. Yöntem

Bu çalışmada geliştirilen yapay zekâ tabanlı veri düzenleme aracı, meta-sezgisel k-nn algoritması ve yapay sinir ağlarının çok katmanlı bir mimari yapıda melezlenmesi ile geliştirilmiştir. İhtiyaç halinde diğer tahmin ve sınıflandırma algoritmaları sisteme dahil edilebilmektedir. Geliştirilmiş olan yapay zekâ tabanlı veri yönetim aracı web sunucusu üzerinde çalışmaktadır. Araştırmacılar web ara yüzü ile etkileşime girip veri setlerini kolay bir şekilde onarabilmektedirler. Şekil 1'de verildiği gibi, yapay zekâ tabanlı veri yönetim aracı veri işleme araçları ile veri seti arasına konumlandırılmaktadır. Yazılım aracı veri seti onarımını yaptıktan sonra, onarılmış veri setini araştırmacılara sunmaktadır.



Şekil 1. Yapay zekâ tabanlı veri yönetim aracının dağılım diyagramı

Yapay zekâ tabanlı veri yönetim aracı Şekil 2'de verildiği gibi 5 adet modülden oluşmaktadır. Bunlar: problem parametrelerini tanımlama modülü, regresyon (tahmin) modülü, gürültülü (eksik veya hatalı) veri tespit modülü, veri onarım modülü ve raporlama modülüdür.



Şekil 2. Veri yönetim süreci yaşam döngüsü

Problem parametrelerini tanımlama modülünde; kullanıcıdan alınan veri seti dosyası (excel tablosu formatında) okunup, problem parametreleri bağımlı ve bağımsız değişkenler olarak etiketlendirilmektedir. Programlama araçlarının sahip olduğu işlevler kullanılarak her bir parametre için veri türü tespiti ve doğrulaması yapılmaktadır (sadece sürekli ya da ayrık sayısal değerli niteliklerden oluşan veri setleri kabul edilecektir). Bu modül bir sonraki regresyon işlemi için probleme ait niteliklerin girişler ve çıkışlar şeklinde etiketlendirmesini sağlayacaktır.

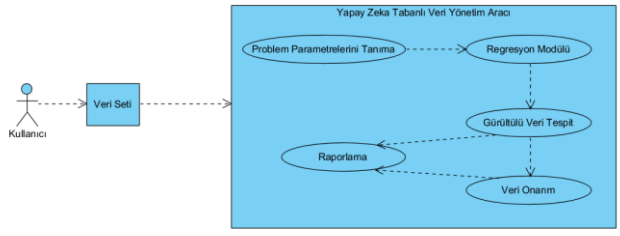
Regresyon (tahmin) modülü: regresyon modülünün giriş verisi, bir önceki modülün çıktısı olan probleme ait etiketlendirilmiş ve veri türleri doğrulanmış niteliklerden oluşan çok boyutlu bir veri setidir. Regresyon modülünün işlevi, problemlere ait bağımlı değişken(ler) ile bağımsız değişkenler arasındaki ilişkileri modellemektir. Bu amaçla modülde, literatürde yer alan iki güçlü yapay zekâ tekniği kullanılarak alternatif modeller oluşturulmaktadır. Bu teknikler yapay sinir ağı (YSA) ve sezgisel k-nn algoritmalarıdır. YSA ile yapılmak istenen, probleme ait bütün nitelikler için bir regresyon modeli yaratmaktır. Çünkü veri setinde herhangi bir veri örneğinin herhangi bir parametresi (niteliği) gürültülü veri içeriyor olabilir. Bu durumda probleme ait bütün parametreler için bir tahmin modeli yaratılması gerekir. Bu aşamada kullanılacak veri seti için birkaç senaryo üzerinden tartışma yapılabilir. İlk senaryo en olumsuz şartlardan biri olsun. Örneğin, veri seti gürültülü veri içeriyorsa bununla oluşturulacak bir regresyon modelinin performansı kötü olur mu? Literatürde YSA algoritması gürültülü veri örneklerine karşı güçlü ve etkili bir teknik olarak bilinmektedir. Bu yönüyle YSA ile oluşturulan modelden tatmin edici bir performans elde etmek mümkündür. Bunun yanında ideal ve olağan senaryo probleme ait kaliteli bir örnek veri setinin (gürültülü verilerden arındırılmış ve probleme ait bilgi alanındaki uzman tarafından doğrulanmış) bu aşamada kullanılmasıdır. Çünkü endüstriyel uygulamalarda ve bilimsel çalışmalarda öncelikle problem uzayını homojen bir şekilde temsil eden veri örneklerinden oluşan bir veri seti hazırlanır. Bu veri seti kullanılarak problem modeli oluşturulur. Daha sonraki aşamalarda ise gerçek bir sistem üzerinden akan veri kullanılarak hedef parametreye yönelik tahmin işlemi gerçekleştirilir. Bu proje çalışmasında geliştirilecek olan regresyon modülünün amacı kaliteli veri üzerinden probleme ait bütün nitelikler için (sadece bağımlı değişken için değil) tahmin modellerini oluşturmak ve bir sonraki aşamayı gerçekleştiren "gürültülü veri tespit modülü" için hazır hale getirmektir. Böylelikle kullanıcıların "gürültülü veri tespit modülüne" yükleyeceği veri setindeki gürültülü verilerin tespit edilmesi ve düzeltilmesi sağlanacaktır. YSA dışında bu modülde parametre tahmini için ayrıca sezgisel k-nn algoritması kullanılmaktadır. Sezgisel k-nn algoritmasının seçilmesinin nedeni YSA'ya alternatif olabilecek nitelikte ve modern bir yapay zekâ tekniği olmasıdır. Regresyon modülünde sezgisel k-nn algoritması ile ilgili yapılacak çalışma bu algoritmanın probleme ve veri setine bağlı olarak değişen parametrelerinin (problem parametrelerinin ağırlık değerleri, k-nn algoritmasının k-değeri, uzaklık

bağıntısı ve oylama yöntemi gibi) ideal değerlerini tespit etmektir [9].

Gürültülü veri tespit modülünde, bir önceki aşamada oluşturulan tahmin modelleri kullanılmaktadır. Bu tahmin modelleri, veri setindeki her bir veri örneği nesnesi için gürültülü değer içeren nitelikleri (bağımlı/bağımsız değişkenleri) tespit etmek için kullanılmaktadırlar. Bu süreçte veri setindeki boş ve dolu hücrelerin kontrolü yapılmaktadır. Boş verilere sahip hücreler işaretlenerek kayıt altına alınmaktadır. Dolu hücrelerdeki veriler için ise tahmin modelleri (YSA ve sezgisel k-nn) ile elde edilen değerler ile bu hücrelerin değerleri karşılaştırılacak ve veri setinde ilgili nitelik için belirlenen standart sapmanın çok üzerinde bir anlamlı farklılık olması durumunda bu hücrenin değeri gürültülü olarak etiketlendirilecektir. Bu süreçte "şüpheli" ve "kesin" şeklinde iki düzeyli bir etiketlendirme yapılmaktadır. Şüpheli durumlarda tahmin sonucunun mu yoksa hücredeki değerin mi kullanılacağına ya da bu veri örneğinin silinme durumuna kullanıcı karar vermektedir. Bu durumlar kullanıcının onayına sunulmaktadır.

Veri onarım modülünde; hata tespit modülünden dönen sonuçlar alınıp, ilgili hücreler üzerinde gerekli işlemler yapılmaktadır. Boş hücrelerin doldurulması için ilgili nitelik için geliştirilen tahmin modeli kullanılarak değer oluşturulmaktadır. Dolu hücrelerdeki veriler için ise "şüpheli" durumda olanlar kullanıcı kararıyla ve "kesin" durumda olanlar da ilgili tahmin modellerinin üreteceği değerler ile düzeltilmektedir.

Hata raporu modülünde; onarımı yapılan hücrenin kayıtları tutulup, kullanıcıya rapor olarak sunulmaktadır. Araç gerçek ortamda çok büyük veriyle karşılaştığında kaç adet değişiklik yapıldı, hangi niteliklerde yüzde kaç problem yaşandı gibi ayrıca istatistiksel olarak raporlama yapılmaktadır.



Şekil 3. Yapay zekâ tabanlı veri yönetim aracının temel öğeleri

3. Deneysel Çalışma

3.1.Veri seti: Enerji Verimliliği Veri Seti (Energy Efficiency Dataset)

3.1.1. Veri Seti Bilgisi:

Ecotect'te simüle edilmiş 12 farklı yapı şeklini kullanarak enerji analizi yapılmaktadır. Veri kümesi, iki gerçek değerli bağımlı değişkeni tahmin etmeyi amaçlayan 768 örnek ve 8 özellikten (bağımsız değişken) oluşur [10].

3.1.2. Özellik Bilgisi:

Veri seti, sekiz bağımsız değişken (X1 ... X8 ile gösterilen özellikler) ve iki bağımlı değişken (Y1 ve Y2 ile gösterilen sonuçlar) içerir. Amaç, iki bağımlı değişkenin her birini tahmin etmek için sekiz bağımsız değişken kullanmaktır [10].

3.1.3. Problem Parametre Karşılıkları:

Tablo 1. Problem parametreleri

X1	*Göreceli Kompaktlık*
X2	*Yüzey Alanı*
X3	*Duvar Alanı*
X4	*Çatı Alanı*
X5	*Genel Yükseklik*
X6	*Yönelim*
X7	*Camlama Alanı*
X8	*Camlama Alanı Dağılımı*
Y1	*Isıtma Yükü*
Y2	*Soğutma Yükü*

3.1.4. Problemin Yapay Sinir Ağları ile Modellenmesi ve Tahmin Modeli Hata Sonuçları

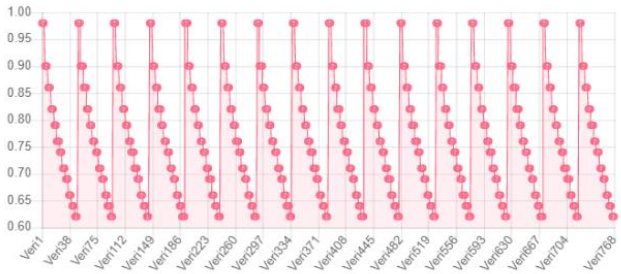
Problem ait veri seti sisteme yüklendikten sonra, sistem veri setine ait her bir problem parametresi (bağımlı veya bağımsız değişken) için bir tahmin modeli (Yapay Sinir Ağı) oluşturmaktadır. Oluşturulan tahmin modeli için MAPE (Ortalama Mutlak Yüzde Hata), MAE (Ortalama Mutlak Hata), MSE (Ortalama Kare Hata), Ortalama Tahmin Hatası ve tahmin modeli sonuçları ile gerçek sonuçların karşılaştırıldığı grafik çıktı olarak verilmektedir.

Tablo 2. Tahmin modelleri hata sonuçları

Problem Parametreleri	*Hatalar*			
	MAPE (%)	*MAE*	*MSE*	*Ortalama Tahmin Hatası*
X1	*0*	*0*	*0*	*0*
X2	*0*	*0*	*0*	*0*
X3	*0*	*0*	*0*	*0*
X4	*0*	*0*	*0*	*0*
X5	*0*	*0*	*0*	*0*
X6	*33.44*	*1*	*1.25*	*0.01*
X7	*3.99*	*0.01*	*0*	*0*
X8	*59.18*	*1.15*	*1.87*	*0.02*
Y1	*0.95*	*0.18*	*0.08*	*0.02*
Y2	*1.12*	*0.24*	*0.12*	*0.01*

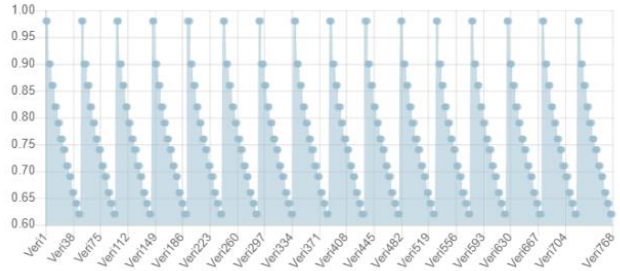
3.1.4.1. X1 parametresine ait tahmin modeli grafikleri

Veri setinde bulunan 768 veri örneği için "Göreceli Kompaktlık (X1)" parametresi veri dağılım grafiği aşağıdaki gibidir.



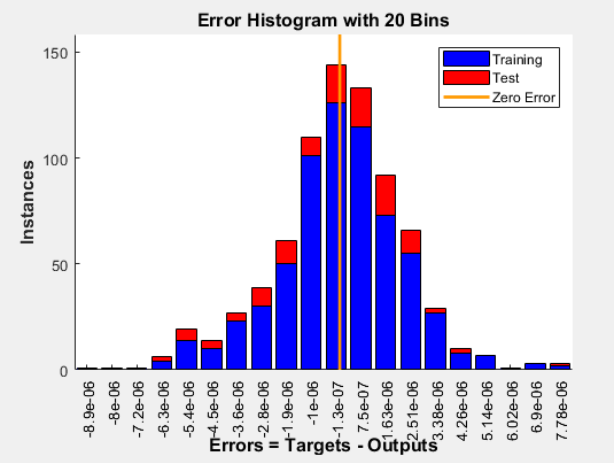
Şekil 4. X1 (1. Problem Parametresi) veri dağılım grafiği

X1 problem parametresini tahmin etmek için diğer 9 nitelikten (X2, X3, X4, X5, X6, X7, X8, Y1, Y2) faydalanılmaktadır. Veri setindeki 768 veri örneği baz alınarak X1 parametresi için bir tahmin modeli (Yapay Sinir Ağı) oluşturulmuştur. Eğitim için kullanılan veri seti, tahmin modeline gönderilmiş ve ele alınan sonuçlar Şekil 5'te ki grafik üzerinde gösterilmiştir.



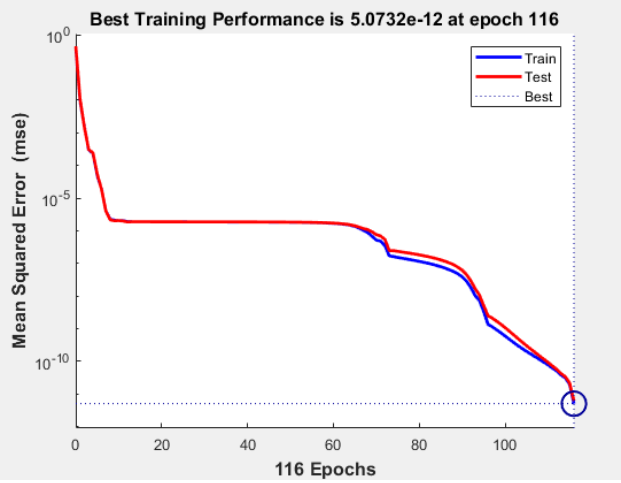
Şekil 5. X1 (1. Problem Parametresi) tahmin modeli sonucu veri dağılım grafiği

X1 problem parametresi için oluşturulan tahmin modelinin eğitimi sırasındaki hata değişimi Şekil 6'da ki histogramda gösterilmektedir.



Şekil 6. X1 (1. Problem Parametresi) tahmin modeli hata histogramı

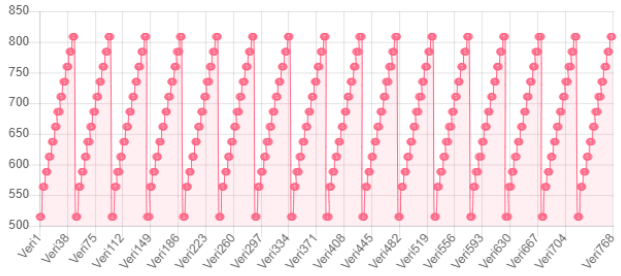
X1 problem parametresi için oluşturulan tahmin modelinin veri seti üzerindeki performansı Şekil 7'de ki grafikte gösterilmektedir.



Şekil 7. X1 (1. Problem Parametresi) tahmin modeli performans grafiği

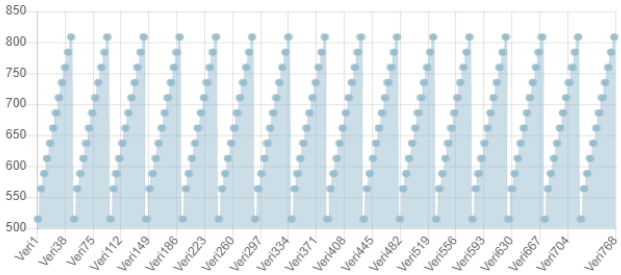
3.1.4.2. X2 parametresine ait tahmin modeli grafikleri

Veri setinde bulunan 768 veri örneği için "Yüzey Alanı (X2)" parametresi veri dağılım grafiği aşağıdaki gibidir.



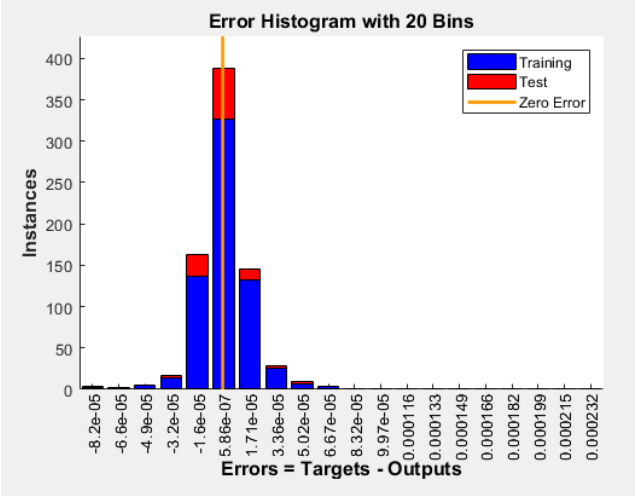
Şekil 8. X2 (2. Problem Parametresi) veri dağılım grafiği

X2 problem parametresini tahmin etmek için diğer 9 nitelikten (X1, X3, X4, X5, X6, X7, X8, Y1, Y2) faydalanılmaktadır. Veri setindeki 768 veri örneği baz alınarak X2 parametresi için bir tahmin modeli (Yapay Sinir Ağı) oluşturulmuştur. Eğitim için kullanılan veri seti, tahmin modeline gönderilmiş ve ele alınan sonuçlar Şekil 9'da ki grafik üzerinde gösterilmiştir.



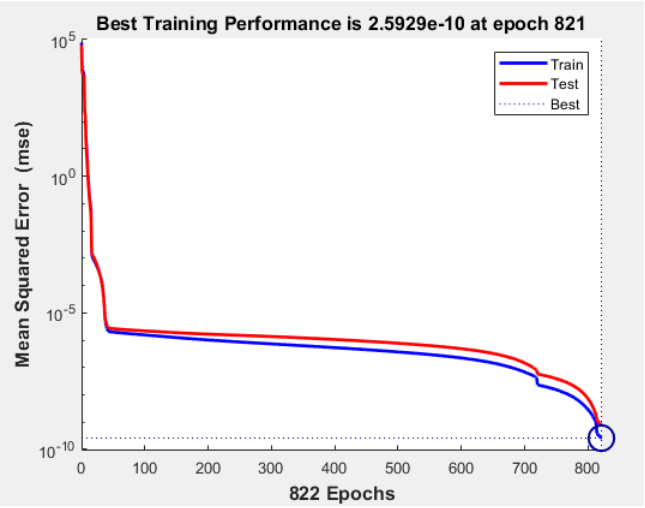
Şekil 9. X2 (2. Problem Parametresi) tahmin modeli sonucu veri dağılım grafiği

X2 problem parametresi için oluşturulan tahmin modelinin eğitimi sırasındaki hata değişimi Şekil 10'da ki histogramda gösterilmektedir.



Şekil 10. X2 (2. Problem Parametresi) tahmin modeli hata histogramı

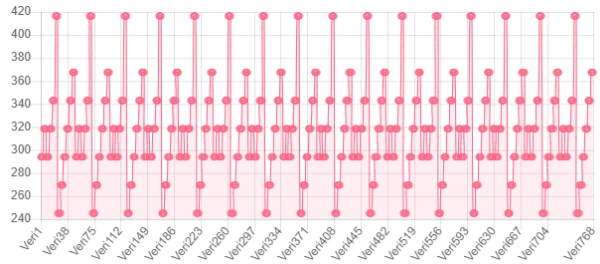
X2 problem parametresi için oluşturulan tahmin modelinin veri seti üzerindeki performansı Şekil 11'de ki grafikte gösterilmektedir.



Şekil 11. X2 (2. Problem Parametresi) tahmin modeli performans grafiği

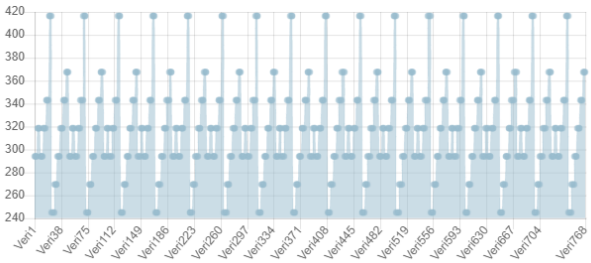
3.1.4.3. X3 parametresine ait tahmin modeli grafikleri

Veri setinde bulunan 768 veri örneği için "Duvar Alanı (X3)" parametresi veri dağılım grafiği aşağıdaki gibidir.



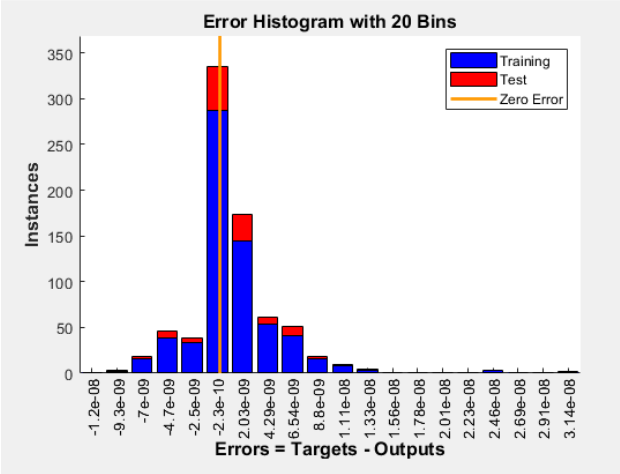
Şekil 12. X3 (3. Problem Parametresi) veri dağılım grafiği

X3 problem parametresini tahmin etmek için diğer 9 nitelikten (X1, X2, X4, X5, X6, X7, X8, Y1, Y2) faydalanılmaktadır. Veri setindeki 768 veri örneği baz alınarak X3 parametresi için bir tahmin modeli (Yapay Sinir Ağı) oluşturulmuştur. Eğitim için kullanılan veri seti, tahmin modeline gönderilmiş ve ele alınan sonuçlar Şekil 13'te ki grafik üzerinde gösterilmiştir.



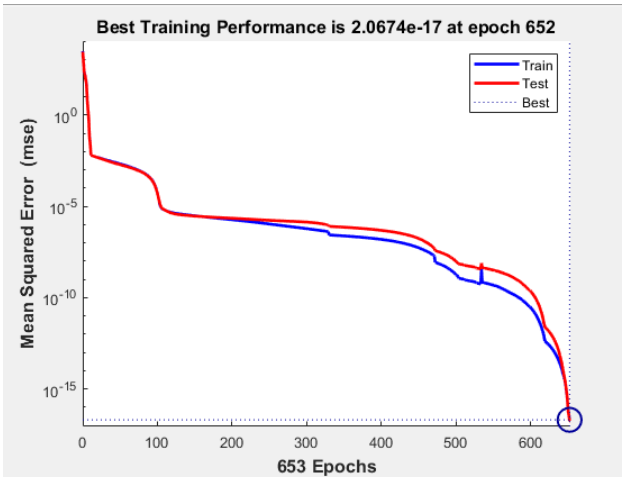
Şekil 13. X3 (3. Problem Parametresi) tahmin modeli sonucu veri dağılım grafiği

X3 problem parametresi için oluşturulan tahmin modelinin eğitimi sırasındaki hata değişimi Şekil 14'te ki histogramda gösterilmektedir.



Şekil 14. X3 (3. Problem Parametresi) tahmin modeli hata histogramı

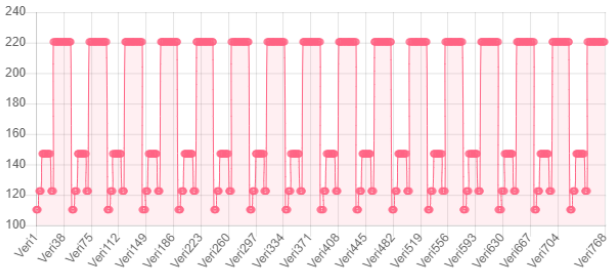
X3 problem parametresi için oluşturulan tahmin modelinin veri seti üzerindeki performansı Şekil 15'te ki grafikte gösterilmektedir.



Şekil 15. X3 (3. Problem Parametresi) tahmin modeli performans grafiği

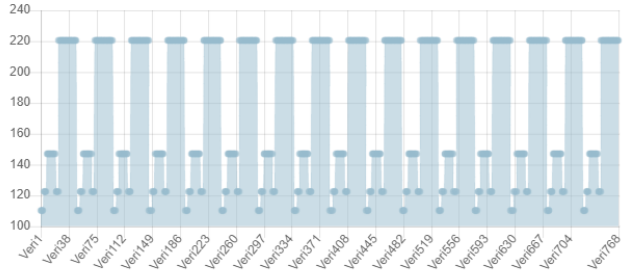
3.1.4.4. X4 parametresine ait tahmin modeli grafikleri

Veri setinde bulunan 768 veri örneği için "Çatı Alanı (X4)" parametresi veri dağılım grafiği aşağıdaki gibidir.



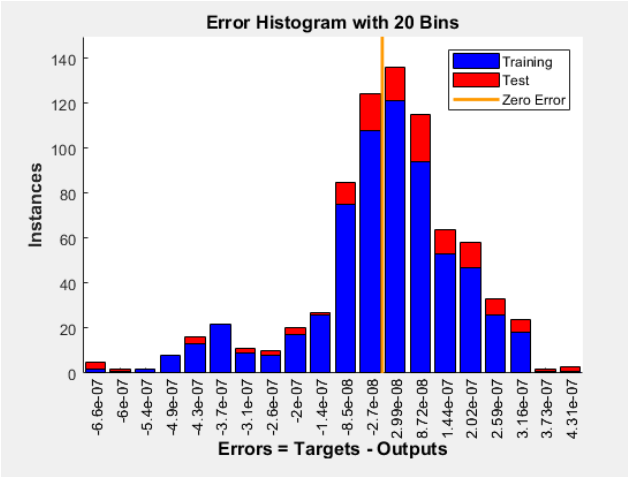
Şekil 16. X4 (4. Problem Parametresi) veri dağılım grafiği

X4 problem parametresini tahmin etmek için diğer 9 nitelikten (X1, X2, X3, X5, X6, X7, X8, Y1, Y2) faydalanılmaktadır. Veri setindeki 768 veri örneği baz alınarak X4 parametresi için bir tahmin modeli (Yapay Sinir Ağı) oluşturulmuştur. Eğitim için kullanılan veri seti, tahmin modeline gönderilmiş ve ele alınan sonuçlar Şekil 17'de ki grafik üzerinde gösterilmiştir.



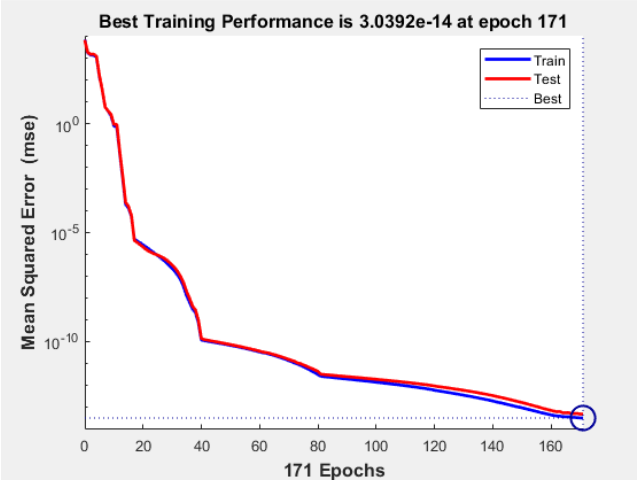
Şekil 17. X4 (4. Problem Parametresi) tahmin modeli sonucu veri dağılım grafiği

X4 problem parametresi için oluşturulan tahmin modelinin eğitimi sırasındaki hata değişimi Şekil 18'te ki histogramda gösterilmektedir.



Şekil 18. X4 (4. Problem Parametresi) tahmin modeli hata histogramı

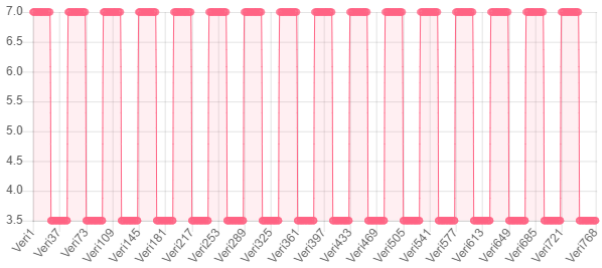
X4 problem parametresi için oluşturulan tahmin modelinin veri seti üzerindeki performansı Şekil 19'da ki grafikte gösterilmektedir.



Şekil 19. X4 (4. Problem Parametresi) tahmin modeli performans grafiği

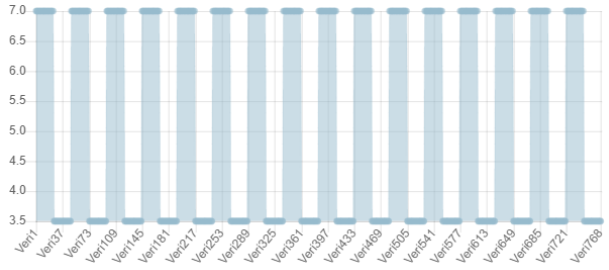
3.1.4.5. X5 parametresine ait tahmin modeli grafikleri

Veri setinde bulunan 768 veri örneği için "Genel Yükseklik (X5)" parametresi veri dağılım grafiği aşağıdaki gibidir.



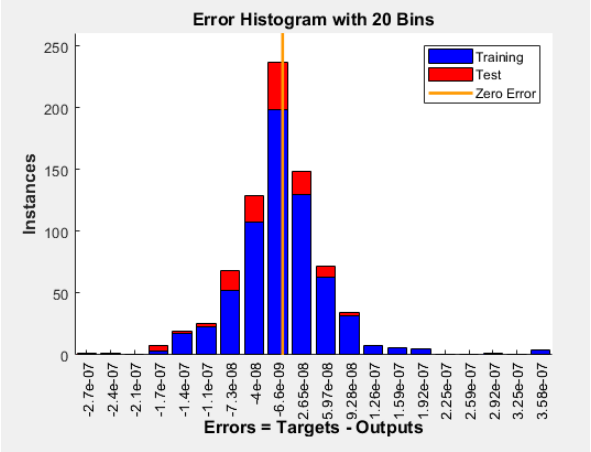
Şekil 20. X5 (5. Problem Parametresi) veri dağılım grafiği

X5 problem parametresini tahmin etmek için diğer 9 nitelikten (X1, X2, X3, X4, X6, X7, X8, Y1, Y2) faydalanılmaktadır. Veri setindeki 768 veri örneği baz alınarak X5 parametresi için bir tahmin modeli (Yapay Sinir Ağı) oluşturulmuştur. Eğitim için kullanılan veri seti, tahmin modeline gönderilmiş ve ele alınan sonuçlar Şekil 21'de ki grafik üzerinde gösterilmiştir.



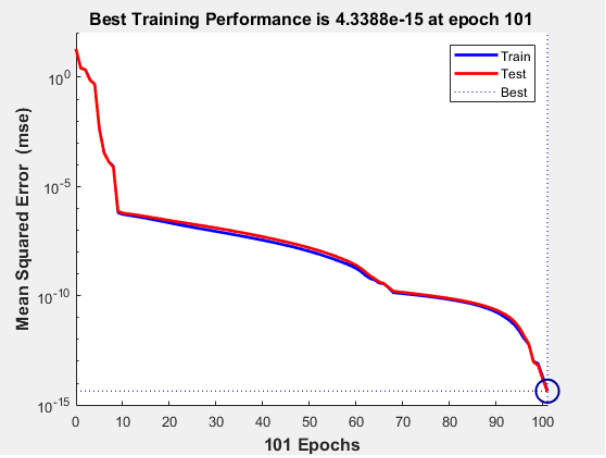
Şekil 21. X5 (5. Problem Parametresi) tahmin modeli sonucu veri dağılım grafiği

X5 problem parametresi için oluşturulan tahmin modelinin eğitimi sırasındaki hata değişimi Şekil 22'de ki histogramda gösterilmektedir.



Şekil 22. X5 (5. Problem Parametresi) tahmin modeli hata histogramı

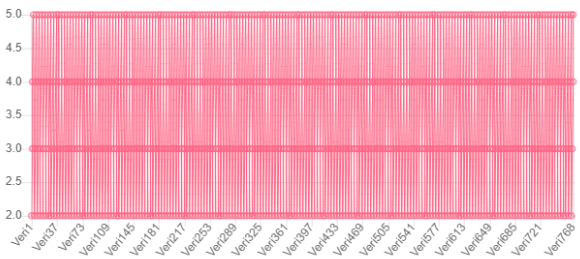
X5 problem parametresi için oluşturulan tahmin modelinin veri seti üzerindeki performansı Şekil 23'da ki grafikte gösterilmektedir.



Şekil 23. X5 (5. Problem Parametresi) tahmin modeli performans grafiği

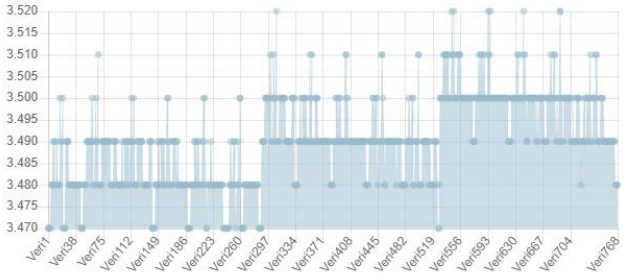
3.1.4.6. X6 parametresine ait tahmin modeli grafikleri

Veri setinde bulunan 768 veri örneği için "Yönelim (X6)" parametresi veri dağılım grafiği aşağıdaki gibidir.



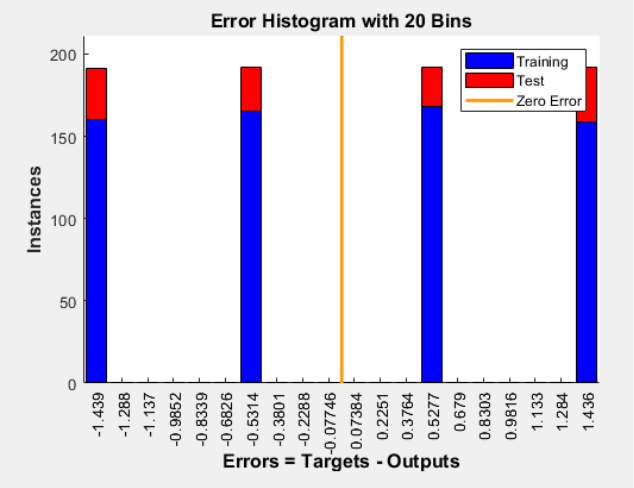
Şekil 24. X6 (6. Problem Parametresi) veri dağılım grafiği

X6 problem parametresini tahmin etmek için diğer 9 nitelikten (X1, X2, X3, X4, X5, X7, X8, Y1, Y2) faydalanılmaktadır. Veri setindeki 768 veri örneği baz alınarak X6 parametresi için bir tahmin modeli (Yapay Sinir Ağı) oluşturulmuştur. Eğitim için kullanılan veri seti, tahmin modeline gönderilmiş ve ele alınan sonuçlar Şekil 25'de ki grafik üzerinde gösterilmiştir.



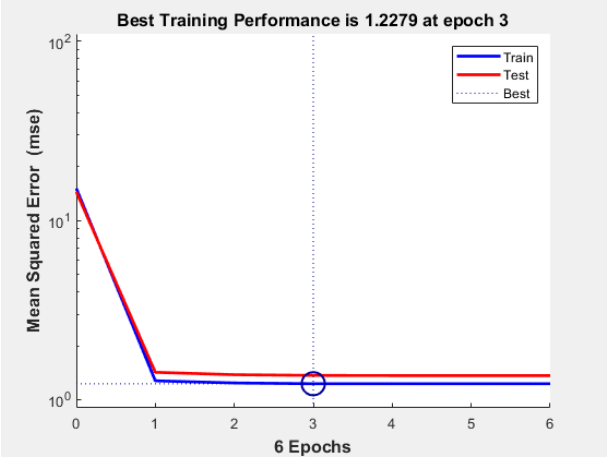
Şekil 25. X6 (6. Problem Parametresi) tahmin modeli sonucu veri dağılım grafiği

X6 problem parametresi için oluşturulan tahmin modelinin eğitimi sırasındaki hata değişimi Şekil 26'da ki histogramda gösterilmektedir.



Şekil 26. X6 (6. Problem Parametresi) tahmin modeli hata histogramı

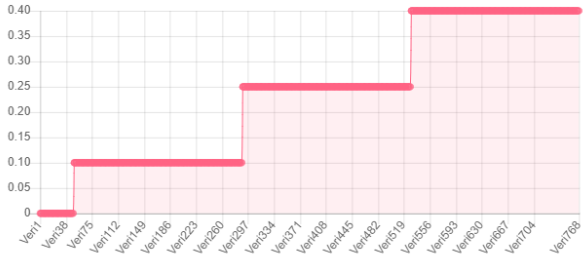
X6 problem parametresi için oluşturulan tahmin modelinin veri seti üzerindeki performansı Şekil 27'de ki grafikte gösterilmektedir.



Şekil 27. X6 (6. Problem Parametresi) tahmin modeli performans grafiği

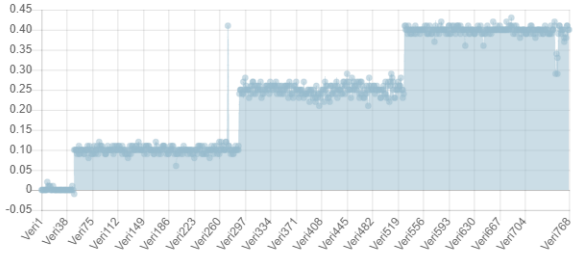
3.1.4.7. X7 parametresine ait tahmin modeli grafikleri

Veri setinde bulunan 768 veri örneği için "Camlama Alanı (X7)" parametresi veri dağılım grafiği aşağıdaki gibidir.



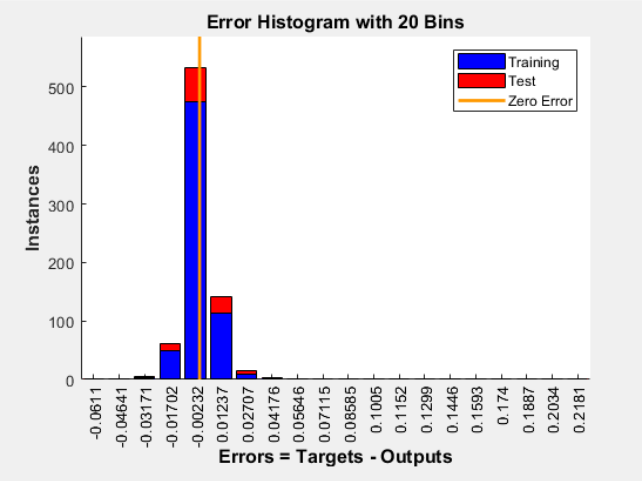
Şekil 28. X7 (7. Problem Parametresi) veri dağılım grafiği

X7 problem parametresini tahmin etmek için diğer 9 nitelikten (X1, X2, X3, X4, X5, X6, X8, Y1, Y2) faydalanılmaktadır. Veri setindeki 768 veri örneği baz alınarak X7 parametresi için bir tahmin modeli (Yapay Sinir Ağı) oluşturulmuştur. Eğitim için kullanılan veri seti, tahmin modeline gönderilmiş ve ele alınan sonuçlar Şekil 29'da ki grafik üzerinde gösterilmiştir.



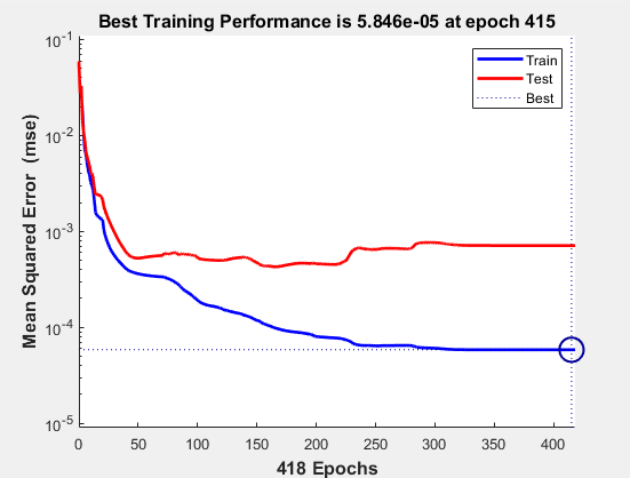
Şekil 29. X7 (7. Problem Parametresi) tahmin modeli sonucu veri dağılım grafiği

X7 problem parametresi için oluşturulan tahmin modelinin eğitimi sırasındaki hata değişimi Şekil 30'da ki histogramda gösterilmektedir.



Şekil 30. X7 (7. Problem Parametresi) tahmin modeli hata histogramı

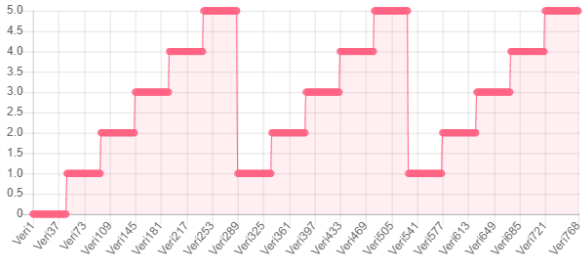
X7 problem parametresi için oluşturulan tahmin modelinin veri seti üzerindeki performansı Şekil 31'de ki grafikte gösterilmektedir.



Şekil 31. X7 (7. Problem Parametresi) tahmin modeli performans grafiği

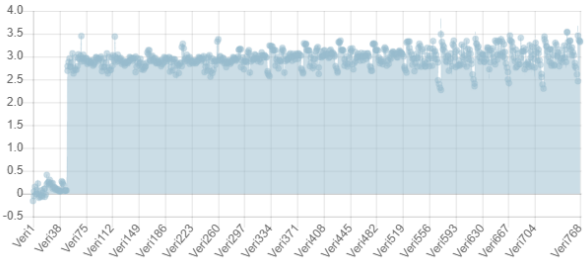
3.1.4.8. X8 parametresine ait tahmin modeli grafikleri

Veri setinde bulunan 768 veri örneği için "Camlama Alanı Dağılımı(X8)" parametresi veri dağılım grafiği aşağıdaki gibidir.



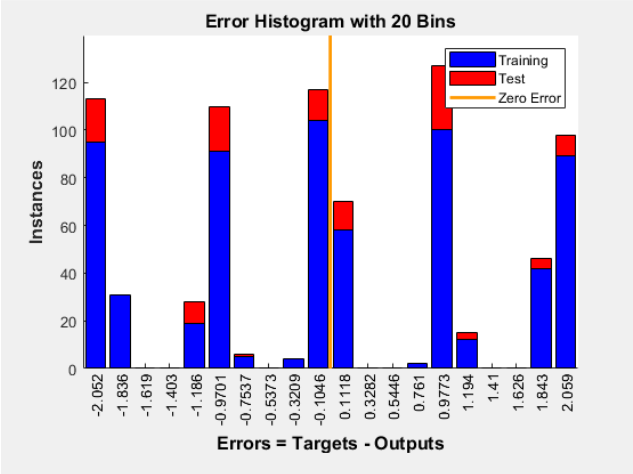
Şekil 32. X8 (8. Problem Parametresi) veri dağılım grafiği

X8 problem parametresini tahmin etmek için diğer 9 nitelikten (X1, X2, X3, X4, X5, X6, X7, Y1, Y2) faydalanılmaktadır. Veri setindeki 768 veri örneği baz alınarak X8 parametresi için bir tahmin modeli (Yapay Sinir Ağı) oluşturulmuştur. Eğitim için kullanılan veri seti, tahmin modeline gönderilmiş ve ele alınan sonuçlar Şekil 33'de ki grafik üzerinde gösterilmiştir.



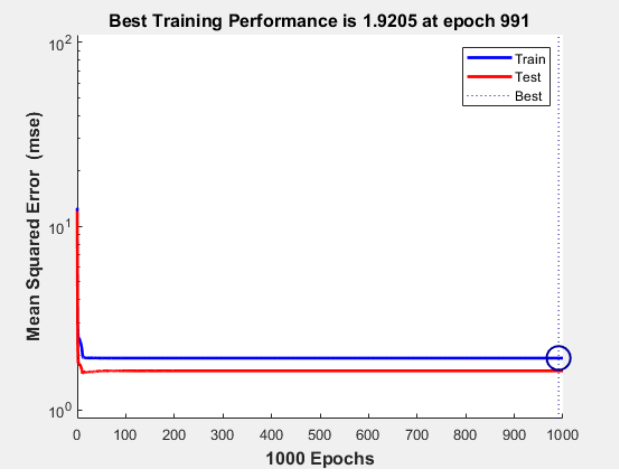
Şekil 33. X8 (8. Problem Parametresi) tahmin modeli sonucu veri dağılım grafiği

X8 problem parametresi için oluşturulan tahmin modelinin eğitimi sırasındaki hata değişimi Şekil 34'de ki histogramda gösterilmektedir.



Şekil 34. X8 (8. Problem Parametresi) tahmin modeli hata histogramı

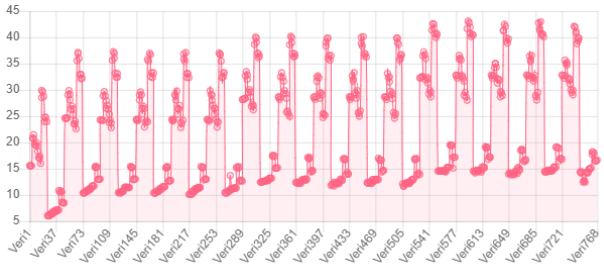
X8 problem parametresi için oluşturulan tahmin modelinin veri seti üzerindeki performansı Şekil 35'de ki grafikte gösterilmektedir.



Şekil 35. X8 (8. Problem Parametresi) tahmin modeli performans grafiği

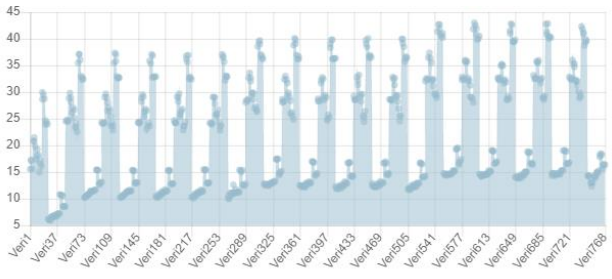
3.1.4.9. Y1 parametresine ait tahmin modeli grafikleri

Veri setinde bulunan 768 veri örneği için "Isıtma Yükü (Y1)" parametresi veri dağılım grafiği aşağıdaki gibidir.



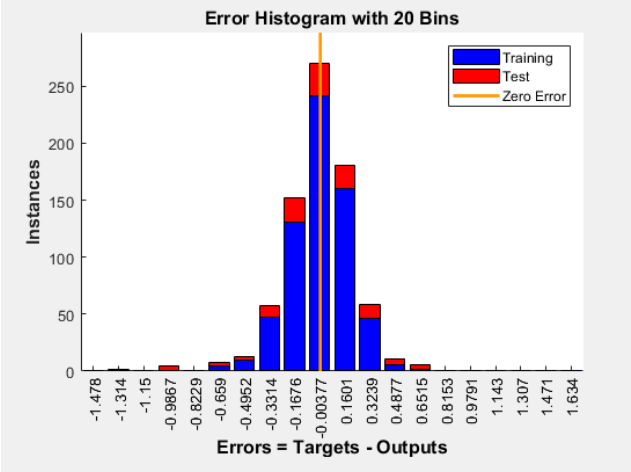
Şekil 36. Y1 (9. Problem Parametresi) veri dağılım grafiği

Y1 problem parametresini tahmin etmek için diğer 9 nitelikten (X1, X2, X3, X4, X5, X6, X7, X8, Y2) faydalanılmaktadır. Veri setindeki 768 veri örneği baz alınarak Y1 parametresi için bir tahmin modeli (Yapay Sinir Ağı) oluşturulmuştur. Eğitim için kullanılan veri seti, tahmin modeline gönderilmiş ve ele alınan sonuçlar Şekil 37'de ki grafik üzerinde gösterilmiştir.



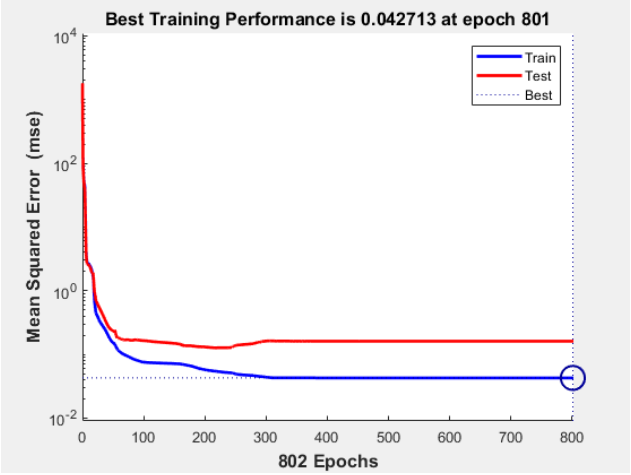
Şekil 37. Y1 (9. Problem Parametresi) tahmin modeli sonucu veri dağılım grafiği

Y1 problem parametresi için oluşturulan tahmin modelinin eğitimi sırasındaki hata değişimi Şekil 38'de ki histogramda gösterilmektedir.



Şekil 38. Y1 (9. Problem Parametresi) tahmin modeli hata histogramı

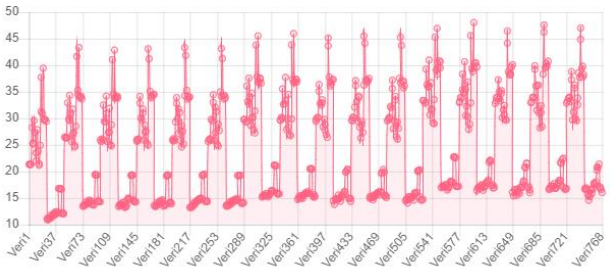
Y1 problem parametresi için oluşturulan tahmin modelinin veri seti üzerindeki performansı Şekil 39'da ki grafikte gösterilmektedir.



Şekil 39. Y1 (9. Problem Parametresi) tahmin modeli performans grafiği

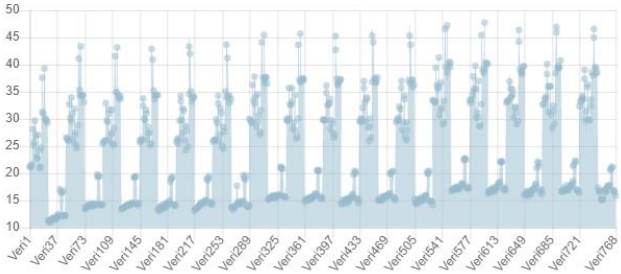
3.1.4.10. Y2 parametresine ait tahmin modeli grafikleri

Veri setinde bulunan 768 veri örneği için "Soğutma Yükü (Y2)" parametresi veri dağılım grafiği aşağıdaki gibidir.



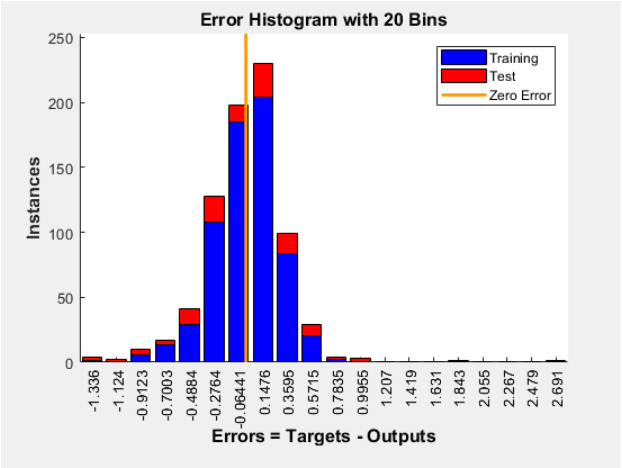
Şekil 40. Y2 (10. Problem Parametresi) veri dağılım grafiği

Y2 problem parametresini tahmin etmek için diğer 9 nitelikten (X1, X2, X3, X4, X5, X6, X7, X8, Y1) faydalanılmaktadır. Veri setindeki 768 veri örneği baz alınarak Y2 parametresi için bir tahmin modeli (Yapay Sinir Ağı) oluşturulmuştur. Eğitim için kullanılan veri seti, tahmin modeline gönderilmiş ve ele alınan sonuçlar Şekil 41'de ki grafik üzerinde gösterilmiştir.



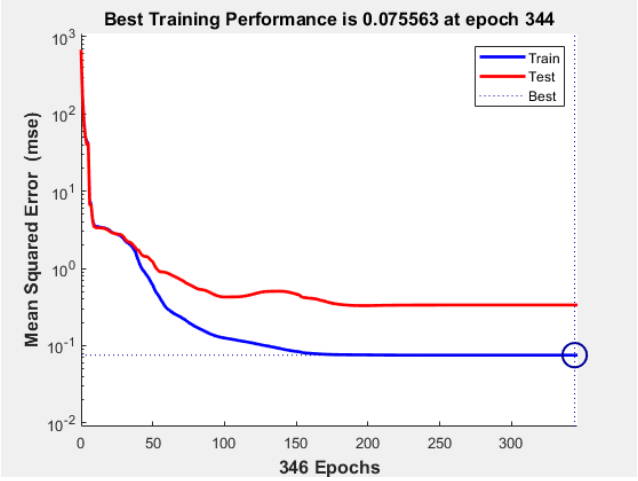
Şekil 41. Y2 (10. Problem Parametresi) tahmin modeli sonucu veri dağılım grafiği

Y2 problem parametresi için oluşturulan tahmin modelinin eğitimi sırasındaki hata değişimi Şekil 42'de ki histogramda gösterilmektedir.



Şekil 42. Y2 (10. Problem Parametresi) tahmin modeli hata histogramı

Y2 problem parametresi için oluşturulan tahmin modelinin veri seti üzerindeki performansı Şekil 43'de ki grafikte gösterilmektedir.



Şekil 43. Y2 (10. Problem Parametresi) tahmin modeli performans grafiği

4. Sonuç

Bu proje önerisinde yapay zekâ tabanlı veri yönetim aracının geliştirilmesi konusunda detaylı bir planlama ve analiz yapılmıştır. Yapay zekâ tabanlı veri yönetim aracının temel öğeleri ve algoritmaları tanımlanmış ve tanıtılmıştır. Bu aracın geliştirilmesine ilişkin yöntem ise detaylı bir şekilde verilmiştir. Geliştirilmesi planlanan aracın literatürde hangi boşluğu dolduracağı yani araştırmacılar açısından önemi açıkça ortaya koyulmuştur. Ortaya koyulan hedefler, geliştirilecek yazılım aracının entegre edileceği sistemleri ve internet üzerinden bağımsız çalışabilen bir uygulamayı işaret etmektedir. Bu yönüyle somut ve ölçülebilir çıktılar tanımlanmıştır. Projenin kabul edilmesi biz proje çalışanlarının yapay zekâ ile veri madenciliği konusunda planladığımız çalışmaları yapmamız açısından son derece önemlidir. Projeyi hayata geçirmemiz, mezuniyet sonrası yapay zekâ alanında faaliyet gösteren yazılım firmalarında işe girmemiz açısından önemli bir referans ve motivasyon kaynağı olacaktır.

Teşekkür

Bu çalışmanın konusunun belirlenmesinde ve hazırlanma sürecinin her aşamasında değerli bilgilerini ve zamanını benden esirgemeyerek her fırsatta çalışmamla yakından ilgilenen, eleştirileriyle yol gösteren danışman hocam Doç. Dr. Hamdi Tolga KAHRAMAN' a teşekkür ve minnetimi özellikle belirtmek istiyorum.

Kaynakça

[1] Deloitte. "Veri Analizi- Veri Kalitesi ve Bütünlüğü". http://www.denetimnet.net/UserFiles/Documents/Makaleler/BT%20Denetim/Veri_Analizi_Veri_Kalitesi_ve_Bütünlüğü.pdf , 13 Ekim 2018.

[2] Çelik, Y., Sezgin, E., "Veri Madenciliğinde Kayıp Veriler İçin Kullanılan Yöntemlerin Karşılaştırılması". Akdeniz Üniversitesi, http://ab.org.tr/ab13/bildiri/184.pdf , 14 Ekim 2018.

[3] Çüm, S., Gelbal, S. (2015). "Kayıp Veriler Yerine Yaklaşık Değer Atamada Kullanılan Farklı Yöntemlerin Model Veri Uyumu Üzerindeki Etkisi", Mehmet Akif Ersoy Üniversitesi Eğitim Fakültesi Dergisi, 35, 87-111.

[4] Çüm, S., Demir, E.K., Gelbal, S., Kışla, T. (2018). "Kayıp Veriler Yerine Yaklaşık Değer Atamak İçin Kullanılan Gelişmiş Yöntemlerin Farklı Koşullar Altında Karşılaştırılması", Mehmet Akif Ersoy Üniversitesi Eğitim Fakültesi Dergisi, 45, 230-249.

[5] Kahraman, H. T., Bayindir, R., & Sagiroglu, S. (2012). A new approach to predict the excitation current and parameter weightings of synchronous machines based on genetic algorithm-based k-NN estimator. Energy Conversion and Management, 64, 129-138.

[6] Karaboga, D., & Akay, B. (2009). A comparative study of artificial bee colony algorithm. Applied mathematics and computation, 214(1), 108-132.

[7] Cheng, M. Y., & Prayogo, D. (2014). Symbiotic organisms search: a new metaheuristic optimization algorithm. Computers & Structures, 139, 98-112.

[8] Holland, J.H., (1975). "Adaptation in natural and artificial systems: An introductory analysis with applications to biology, control, and artificial intelligence". Q. Rev. Biol. 1, 211. http://dx.doi.org/10.1086/418447.

[9] Kahraman, H. T. (2016). A novel and powerful hybrid classifier method: Development and testing of heuristic k-nn algorithm with fuzzy distance metric. Data & Knowledge Engineering, 103, 44-59.

[10] A. Tsanas, A. Xifara: 'Accurate quantitative estimation of energy performance of residential buildings using statistical machine learning tools', Energy and Buildings, Vol. 49, pp. 560-567, 2012.

[11] Arslan, F.,(2019),'Yapay Zekâ Tabanlı Büyük Veri Yönetim Aracının Tasarımı ve Uygulaması', Karadeniz Teknik Üniversitesi Lisans Bitirme Tezi.