VERİ MADENCİLİĞİ VE ÇİMENTO SEKTÖRÜNDE BİR UYGULAMA

Adil Baykasoğlu

Gaziantep Üniversitesi, Endüstri Mühendisliği Bölümü, 27310 Gaziantep
Tel-Faks: 0342 3604383, E-posta: baykasoglu@gantep.edu.tr

Özet

Veri tabanlarından daha önce bilinmeyen faydalı bilgilerin çıkarımı diye isimlendirebileceğimiz veri madenciliği günümüz veri yoğun dünyasının vazgeçilmez uygulamalarından birisi olma yolundadır. Bu çalışmada veri madenciliği tanıtılarak, eğitim, sağlık, bankacılık, perakendecilik vb alanlarda yapılan başarılı uygulamalarından bahsedilerek çimento sektöründe gerçekleştirilen bir çalışması tanıtılacaktır.

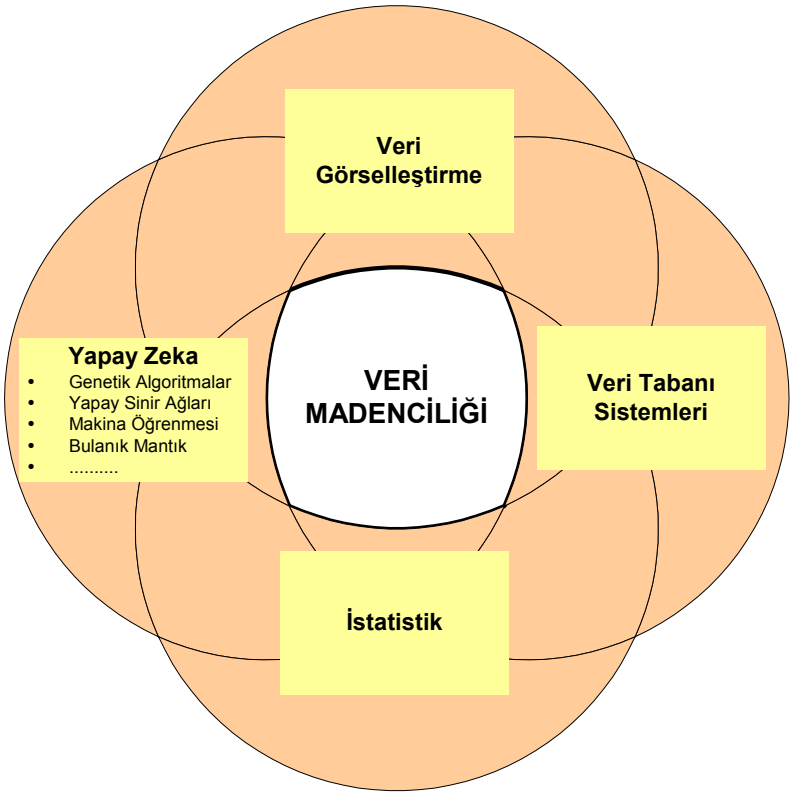
Anahtar Kelimeler: Veri madenciliği, yapay zeka, veri tabanları

1. GİRİŞ

Veri madenciliği, eldeki verilerden üstü kapalı, çok net olmayan, önceden bilinmeyen ancak potansiyel olarak kullanışlı bilginin çıkarılmasıdır. Bu da; kümeleme, veri özetleme, değişikliklerin analizi, sapmaların tespiti gibi belirli sayıda teknik yaklaşımları içerir. Başka bir deyişle, veri madenciliği, verilerin içerisindeki desenlerin, ilişkilerin, değişimlerin, düzensizliklerin, kuralların ve istatistiksel olarak önemli olan yapıların yarı otomatik olarak keşfedilmesidir.

Temel olarak veri madenciliği, geniş veri setleri arasındaki desenlerin yada düzenin, verinin analizi ve yazılım tekniklerinin kullanılması ile ilgilidir. Veriler arasındaki ilişkiyi, kuralları ve özellikleri belirlemekten bilgisayar sorumludur. Amaç, daha önceden fark edilmemiş veri desenlerini tespit edebilmektir.

Veri madenciliğini istatistiksel bir yöntemler serisi olarak görmek mümkün olabilir. Ancak veri madenciliği, geleneksel istatistikten birkaç yönde farklılık gösterir. Veri madenciliğinde amaç, kolaylıkla mantıksal kurallara yada görsel sunumlara çevrilebilecek nitel modellerin çıkarılmasıdır. Şekil 1'de de gösterildiği gibi veri madenciliği sahası, istatistik, yapay zeka, veri tabanları ve veri görselleştirme gibi alanlar ile yakından ilişkilidir.



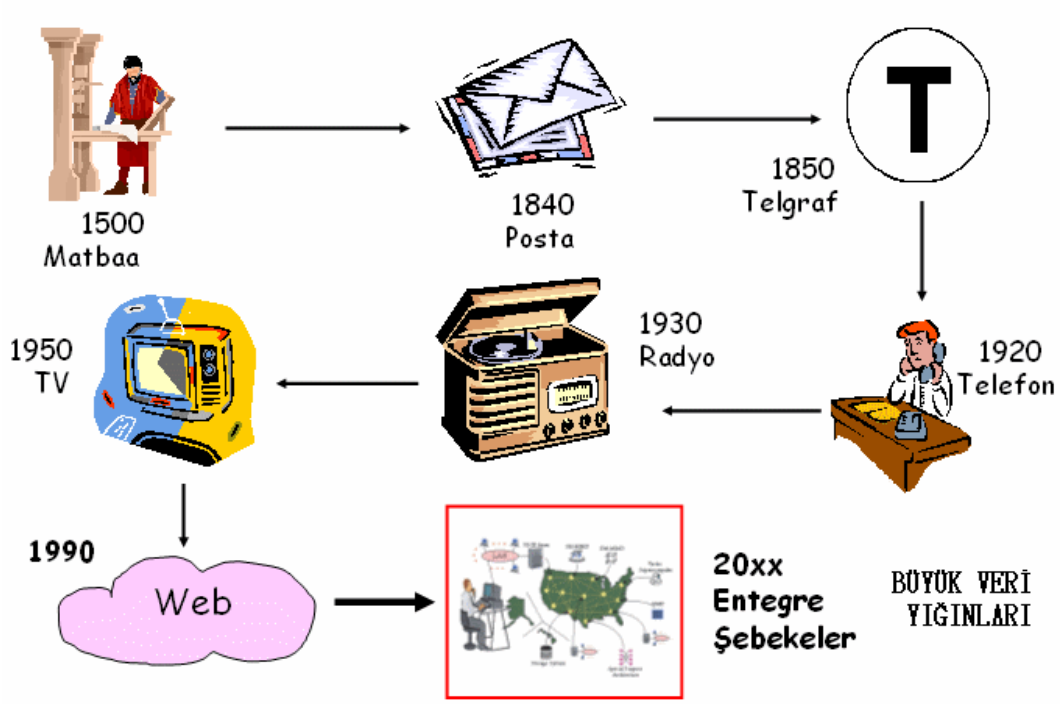
Şekil 1. Veri madenciliği ilgili alanları

Veri madenciliği konusunda bahsi geçen geniş verideki *geniş* kelimesi, tek bir iş istasyonunun belleğine sığamayacak kadar büyük veri kümelerini ifade etmektedir. Yüksek hacimli veri ise, tek bir iş istasyonundaki yada bir grup iş istasyonundaki disklere sığamayacak kadar fazla veri anlamındadır. Dağıtık veri ise, farklı coğrafi konumlarda bulunan verileri anlatır. Araştırmacıların, geniş, çok hacimli ve dağınık veri setleri üzerinde yapmış oldukları çalışmalar sonucu aşağıdaki sonuçlara varılmıştır [1].

- Veri madenciliği ve bilgi keşfi, özellikle elektronik ticaret, bilim, tıp, iş ve eğitim alanlarındaki uygulamalarda yeni ve temel bir araştırma sahası olarak ortaya çıkmaya başlamıştır. Veri madenciliği, eldeki yapısız veriden, anlamlı ve kullanışlı bilgiyi çıkarmaya yarayacak tümevarım işlemlerini formüle analiz etmeye ve uygulamaya yönelik çalışmaların bütününü içerir. Geniş veri kümelerinden desenleri, değişiklikleri, düzensizlikleri ve ilişkileri çıkarmakta kullanılır. Bu sayede, web üzerinde filtrelemeler, DNA sıraları içerisinde genlerin tespiti, ekonomideki eğilim ve düzensizliklerin tespiti, elektronik alışveriş yapan müşterilerin alışkanlıkları gibi karar verme mekanizmaları için önemli bulgular elde edilebilir.
- Sayısal verinin miktarı, son 10 yılda bir patlama yaşayarak tahminlerin dışında bir artış göstermiştir. Buna karşılık, bilim adamlarının, mühendislerin ve analistlerin sayısı değişmemektedir. Bu orantısızlığı gidermek için yeni araştırma problemlerinin çözümleri birkaç gruba ayrılabilir:
  - Geniş hacimli ve çok boyutlu veri madenciliği için yeni algoritma ve sistemlerin geliştirilmesi,
  - Yeni veri tiplerinin madenciliği için yeni algoritma, teknik ve sistemlerin geliştirilmesi,
  - Dağıtık veri madenciliği için algoritma, protokol ve altyapıların geliştirilmesi,
  - Mevcut veri madenciliği sistemlerinin kullanımının ilerletilip geliştirilmesi,
  - Veri madenciliği için özel gizlilik ve güvenlik modellerinin geliştirilmesi.
- Tüm bu uğraşların başarıya ulaşması ve sonuç verebilmesi için çok disiplinli ve disiplinler arası çalışan iş sahalarının desteği gereklidir.

- İlgili sistemlerin, ölçülmüş altyapıların ve test ortamlarının oluşturulmasını gerektiren önemli deneysel bileşenlerin gerçekleştirilmesi gerekir.

Yukarıda da özetlenmeye çalışıldığı gibi veri madenciliği çalışmaları günümüz bilgi toplumunda kritik bir alan olmaya başlamıştır. Bilişim, Internet ve medya teknolojilerindeki olağan üstü gelişmeler bizleri bir veri okyanusu ile karşı karşıya bırakmıştır (Bkz Şekil 2.) Bu veri okyanusundan bilgiye ulaşmak için bir başka ifade ile balık tutmak için özellikle Avrupa ve ABD de veri madenciliği konusunda birçok araştırma gurubu kurulmuş ve kurulmaktadır. 22 Mayıs 2000 tarihli *Time* dergisinde yer alan bir yazıda veri madenciliği en sıcak on iş alanından birisi olarak gösterilmiştir. *MIT's Magazine of Technology Review* dergisinin Oçak-Şubat 2001 nolu sayısı ise veri madenciliğini teknoloji alanında ortaya çıkan 10 yeni alandan biri olarak işaret etmiştir. *Ekonomist Online* ise 2004 yılında veri madenciliğini yarının 12 harikasından biri olarak göstermiştir.



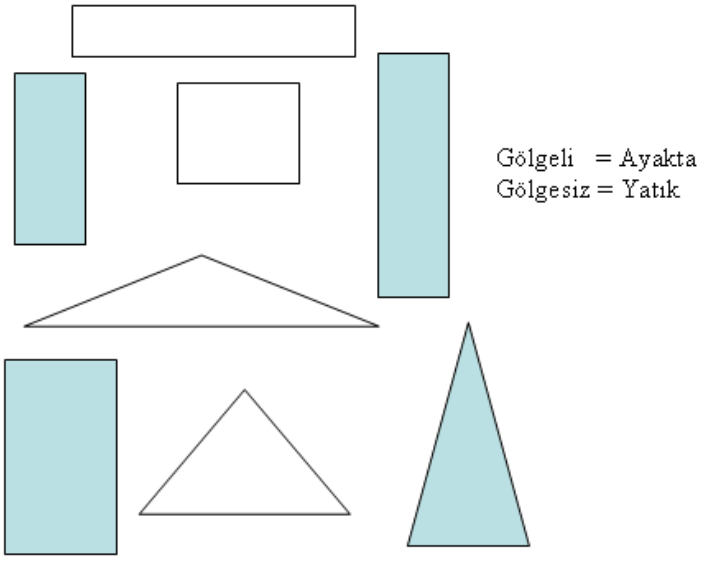
Şekil 2. Gelişen bilişim teknolojisi ve veri oluşumu

Veri madenciliği ile ilgili literatürde çok değişik uygulamalara rahatlıkla ulaşmak mümkündür. Tıp ve mühendislik alanları diğer alanlara göre uygulamaların daha az olduğu alanlar olarak gözükmektedir. Sosyal bilimlerde daha yoğun uygulamalar mevcuttur (e-ticaret, kredilendirme, müşteri ilişkilileri, iş yeri tasarımı, yer secimi vb.). Bu çalışmada veri madenciliği mühendislik alanındaki bir probleme uygulanmıştır. Problem çimento sektöründe çok önemli bir konu olan çimentonun basma dayanımının tahmin edilmesi ile ilgilidir. Üretilen çimentodan elde edilen numunelerin 28 gün bekletilerek yapılan mukavemet deneyi çimentonun kalitesi ile ilgili en önemli parametrelerden birisidir. Ancak deney sonuçlarını elde etmek için 28 günlük bir beklemenin gerekli olması pratik üretim koşulları için uygun olmamaktadır. Bu nedenle çimento mukavemetinin önceden tahmin edilebilmesi önemlidir.

Makalede önce veri madenciliği teknikleri ve uygulama alanlarından bahsedilecek daha sonra yapılan uygulama çalışması tanıtılacaktır.

2. VERİ MADENCİLİĞİ YÖNTEM ve TEKNİKLERİ

Yöntem ve tekniklere girmeden önce basit bir kural çıkarma örneği ile veri madenciliğini tarif etmeye çalışalım. Şekil 3 de bazı geometrik şekiller veri olarak verilmiş olsun.



Şekil 3. Şekiller problemi

Araştırma sorumuz veya amacımız bu verinin ne tür bir bilgiyi içerdiğini araştırmak olsun. Şekillerin geometrik ölçüleri Tablo 1 de gösterildiği gibi elde edilip herhangi bir kümeleme algoritması uygulandığında aşağıdaki kurallar kolayca elde edilecektir.

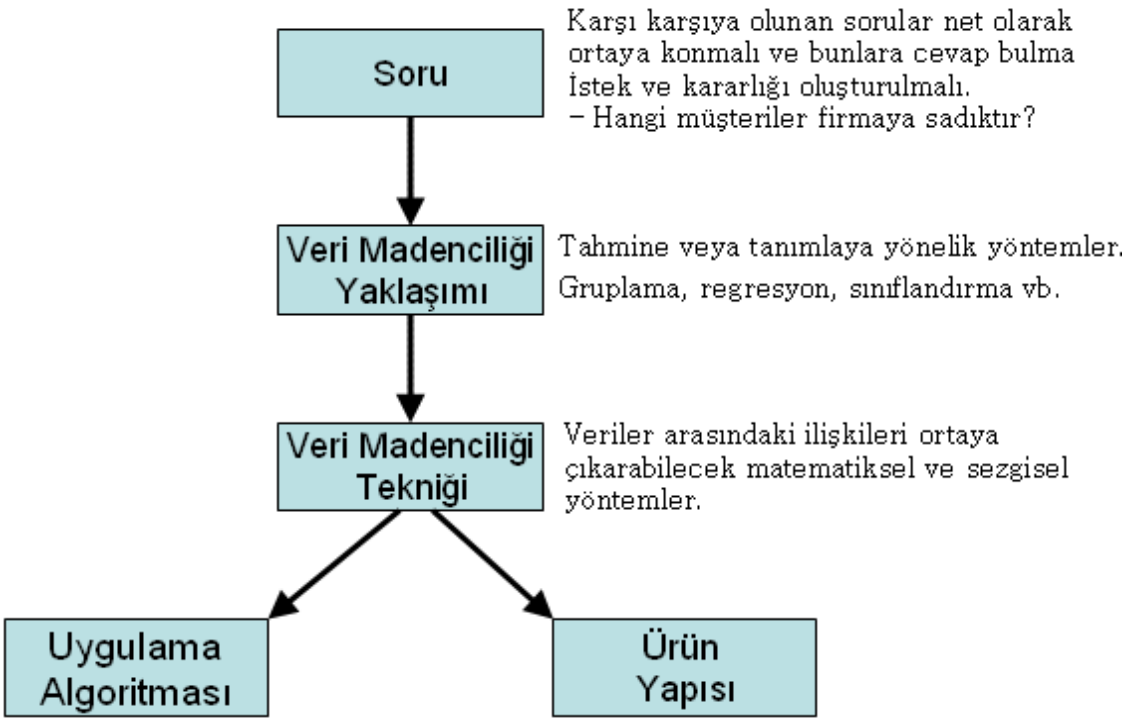
Tablo 1. Şekiller problem verisi

Genişlik	Yükseklik	Kenar	Küme
2	4	4	ayakta
3	6	4	ayakta
4	3	4	yatık
7	8	3	ayakta
7	6	3	yatık
2	9	4	ayakta
9	1	4	yatık
10	2	3	yatık

Kümeleme Kuralları:
IF genişlik ≥ 3.5 AND yükseklik < 7.0 THEN yatık
IF yükseklik ≥ 3.5 THEN ayakta

İlişkisel Kurallar:
IF genişlik > yükseklik THEN yatık
IF yükseklik > genişlik THEN ayakta

Veri madenciliği yöntemi Şekil 4 de tarif edildiği gibi araştırma sorusunun net olarak ortaya konması ile başlayan bir süreçtir. İkinci adım bulunmaya çalışılan cevapa ve eldeki veriye uygun bir veri madenciliği yaklaşımın tespit edilmesidir. Üçüncü adımda seçilen yaklaşım kümesinden bir veya birkaç tekniğin veriye uygulanarak bilginin keşfi aşamasıdır. Bilgi keşfi süreci Şekil 5 de ayrıntılı bir biçimde gösterilmiştir.

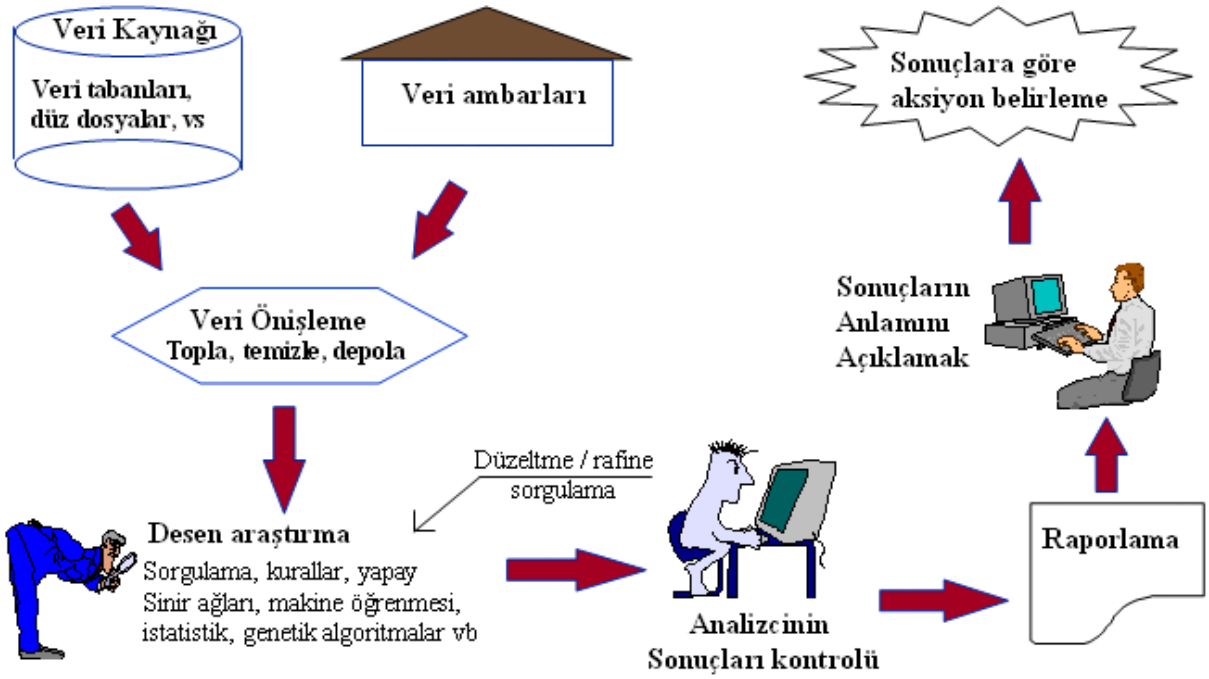


Şekil 4. Veri madenciliğin yapısı

Veri madenciğinde kullanılmak üzere birçok yöntem ve algoritma geliştirilmiştir ve geliştirilmeye de devam edilmektedir. Bu yöntemlerden en çok kullanılanlarını aşağıdaki gibi sınıflandırmak mümkündür [2].

İstatistiksel Yöntemler: Veri madenciliği temel olarak ileri bir istatistik uygulamasıdır. Verilen bir örnek kümesine bir kestirici oturtmayı amaçlar. İstatistik literatüründe son elli yılda bu amaç için birçok teknik geliştirilmiştir. Bu tekniklerin en önemlileri çok boyutlu analiz başlığı altında toplanır ve genelde verinin parametrik bir modelden geldiğini varsayar. Bu varsayım altında sınıflandırma, regresyon, öbekleme, boyut azaltma, hipotez testi, varyans analizi, bağıntı kurma için teknikler istatistikte uzun yıllardır kullanılmaktadır.

Örnek Temelli Yöntemler: Bu yöntem istatistikçiler tarafından 1950'li yıllarda önerilmiş olmasına rağmen o yıllarda gerektirdiği hesaplama ve bellek yüzünden kullanılamamış ama günümüzde bilgisayarların kapasitelerinin artmasıyla, özellikle de çok işlemcili sistemlerin yaygınlaşmasıyla, kullanılabilir olmuştur. Bu yönteme en iyi örnek en yakın k komşu algoritması ve vaka temelli muhakemedir.

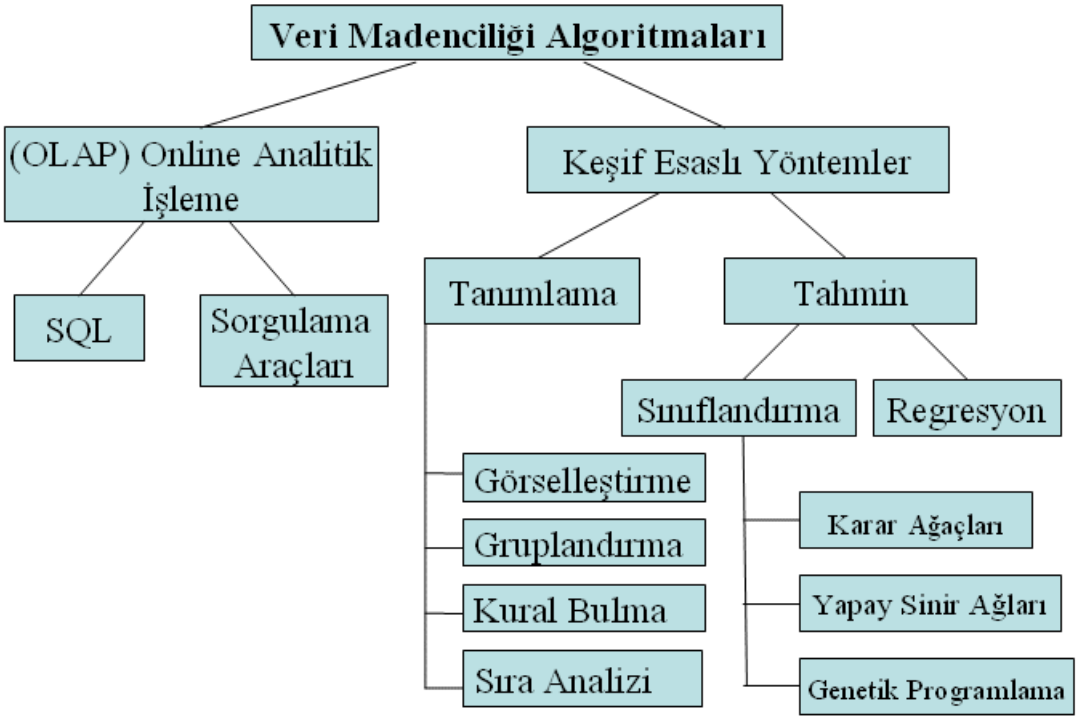


Şekil 5. Veri madenciliğin bilgi keşfi prosesi

Yapay Zeka Temelli Yöntemler: Bu yöntemler veri madenciliğinin beklide en etkili yöntemleri arasında yer alan yapay sinir ağları, genetik algoritmalar, genetik programlama, bulanık mantık gibi modern teknikleri içerir. Yapılan pek çok çalışma bu tekniklerin özellikle doğrusal olmayan, bulanık ve değişken yapıdaki veri uzaylarında klasik istatistik temelli yöntemlere göre çok daha başarılı sonuçlar verdiğini göstermektedir.

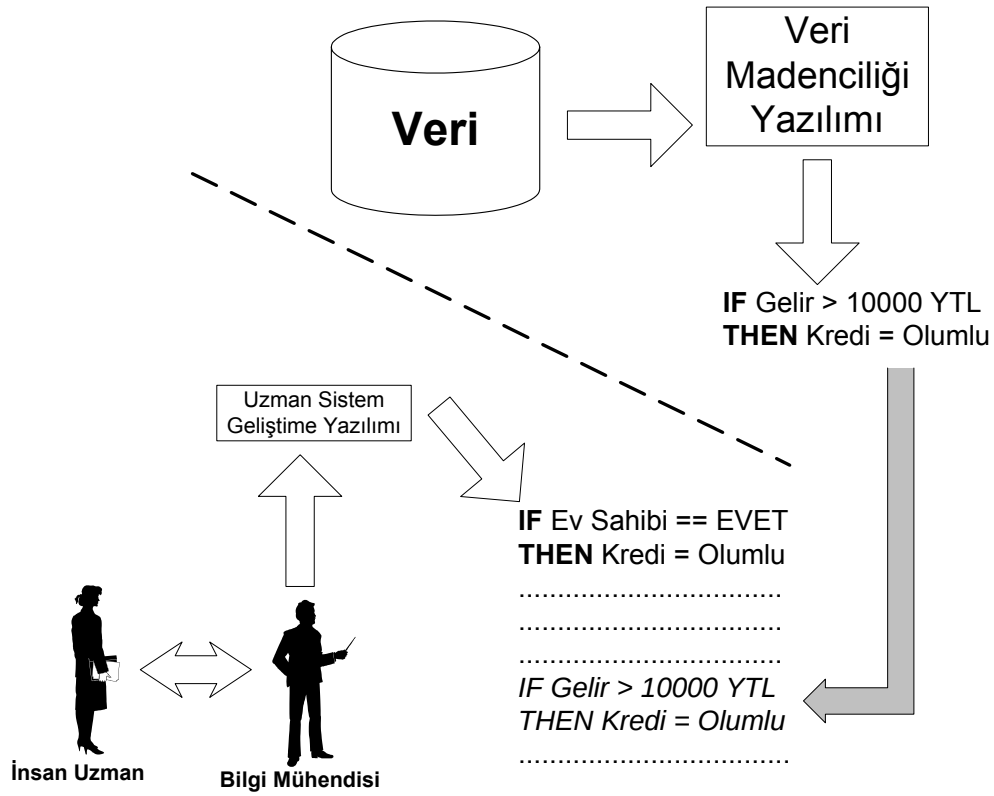
Karar Ağaçları: Karar ağaçları esas itibari ile kural çıkarma algoritmalarıdır. Bu algoritmalar bir veri kümesinden kullanıcıların çok kolay anlayabileceği (IF-THEN) türündeki kuralları bir ağaç yapısında türetebiliriler. Karar ağaçları veri madenciliği uygulamalarında en çok kullanılan ve en fazla yazılımın bulunduğu algoritmalar kümesini oluşturur.

Veri madenciliğinde kullanılan algoritmaları yöntem bağımsız olarak sınıflandıracak olursak Şekil 6'daki gibi bir sınıflandırma yapılabilir.



Şekil 6. Veri madenciliği algoritmaları

Veri madenciliğinin benzerlik gösterdiği alanlardan bir tanesi de uzman sistemler alanıdır. Ancak veri madenciliği bir uzman sistem yaklaşımı değildir. Her iki yöntem arasındaki fark Şekil 7 de grafiksel olarak gösterilmiştir. Fakat heyecan verici olan veri madenciliğinin uzman sistemleri besleyebilecek bir yaklaşım olmasıdır. Uzman sistemlerim en önemli sorunu olan kural tabanını oluşturacak olan kuralların elde edilmesi veri yoğun ortamlarda veri madenciliği yaklaşımları ile desteklenebilir. Örneğin kanser teşhisini amaçlayan bir uzman sistem geliştirilmesi amaçlanıyorsa uzman doktorlardan elde edilecek olan kurallar kanser veri tabanından elde edilebilecek verilerle çok daha zenginleştirilebilir. Bu şekilde çok daha güvenilir uzman sistemlerin oluşturulması mümkün olabilir.



Şekil 7. Veri madenciliği ve uzman sistemler

3. VERİ MADENCİLİĞİNİN BAZI UYGULAMA ALANLARI

Veri madenciliği verinin yoğun olarak üretildiği hemen her ortamda uygulama kullanım alanı bulabilir ve bulmuştur. Bazı uygulama alanları şu şekilde özetlenebilir.

- *Bilimsel ve mühendislik verileri* : Günümüzde laboratuar veya bilgisayar ortamında sistemlerin benzetimi ve analizi sürecinde yüksek miktarda bilimsel veri üretilmektedir. Elde edilen bu verilerin anlamlandırılması için veri madenciliği çok uygun bir platform sağlar. Çimento deneylerinde elde edilen verilerden mukavemet analizi, üretim sistemlerinin benzetiminden elde edilen verilerden sistem performansını etkileyen faktörlerin ve kuralların çıkarılması, deprem verilerinin analizi ile deprem ve etkilerinin tahmini, kalite kontrol uygulamaları gibi pek çok uygulama örnek olarak verilebilir.
- *Sağlık verileri* : Veri madenciliğinin en umut verici uygulama alanlarından bir tanesi de tıp ve sağlık alanıdır. Özellikle tarama testlerinden elde edilen verileri kullanarak çeşitli kanserlerin ön tanısı, kalp verilerini kullanarak kalp krizi riskinin tespiti, acil servislerde hasta semptomlarına göre risk ve önceliklerin tespiti gibi çok geniş bir uygulama sahası söz konusudur. Vysis, ilaç sanayisi için protein analizini yapay sinir ağlar algoritmasını kullanarak yapmaktadır. Rochester Kanser Merkezi bölümü araştırmalarında KnowledgeSEEKER adlı karar ağacı tekniğini kullanır. California Hastenesi veriden bilgi üretmek için "Information Discovery" adlı ürünü kullanmaktadır. Hastanede çalışan bir doktor, bu program sayesinde hastalarını hiçbir fiziksel teste tabii tutmadan teşhis koyabildiğini açıklamıştır [3].
- *İş verileri* : iş süreçleri sırasında geniş çaplarda veri üretirler. Bu veriyi karar verme mekanizmalarında efektif olarak kullanılabilir. Personel veri tabanlarında yapılabilecek analizler ile performansa etki eden faktörlerin tespiti ve yeni personel seçiminde kullanılabilecek değişik kuralar türetilebilir. Müşteri veri tabanlarının analizi ile reklam ve promosyon ile ilgili pek çok faydalı bilgiye ulaşmak mümkündür.

- *Perakendecilik – marketçilik verileri* : Bu alanda en çok başvurulan veri madenciliği yaklaşımı sepet analizidir. Sepet analizinde amaç alınanlar ürünler arasındaki ilişkileri bulmaktır. Bu ilişkilerin bilinmesi işletmenin kârını arttırmak için kullanılabilir. Eğer A ürününü alanların B ürününü de çok yüksek olasılıkla aldıklarını biliyorsanız ve eğer bir müşteri A ürününü alıyor ama B ürününü almıyorsa o potansiyel bir B müşterisidir. O halde bu ürünler tüketicinin dikkatini çekmek için bir arada sergilenebilir.
- *Bankacılık, finans ve borsa verileri* : Bankacılık sektöründe kredi ve kredi kartı sahtekarlığı tahminlerinde, risk değerlendirmede, müşteri eğilim analizlerinde, kar analizi gibi alanlarda veri madenciliği kullanılır. Finans ve borsa kuruluşları ise stok fiyat tahminlerinde, gümrük ölçümleme, portföy yönetimi gibi alanlarda veri madenciliği yöntemlerini kullanabilirler.
- *Eğitim sektörü verileri* : Öğrenci işlerinde veriler analiz edilerek öğrencilerin başarı ve başarısızlık nedenleri, başarının arttırılması için hangi konulara ağırlık verilmesi gerektiği, üniversite giriş puanları ile okul başarısı arasında bir ilişkinin var olup olmadığı gibi onlarca sorunun cevabı bulunarak eğitim kalitesi ve performansı arttırılabilir.
- *Internet (Web) verileri* : Internet ve web üzerindeki veriler hem hacim hem de karmaşıklık olarak hızla artmaktadır. Web madenciliği özetle internetten faydalı bilginin keşfi olarak tanımlanabilir. Web veri madenciliği birçok web sunucusu veya online servisten kullanıcı erişim desenlerinin analiz ve keşfi için kullanılır[4].Örneğin internet üzerinden kitap satan Amazon şirketi (http://www.amazon.com) *BookMatcher* adlı programıyla müşterilerine okudukları ve sevdikleri kitaplara göre satın almaları için kitap tavsiye etmektedir.
- *Doküman verileri* : Doküman veri madenciliğinde (text mining) ana amaç dokümanlar arasında ayrıca elle bir tasnif gerekmeden benzerlik hesaplayabilmektir. Bu genelde otomatik olarak çıkarılan anahtar sözcüklerin tekrar sayısı sayesinde yapılır. Polis kayıtlarında mevcut rapora benzer kaç adet ve hangi raporlar var. Ürün tasarım dokümanları ve internet dokümanları arasında mevcut tasarım için kullanılabilecek ne tür dosyalar var gibi sorulara yanıt bulunabilir.

Bunların dışında da veri madenciliğinin faydalı olabileceği ve kullanılabileceği alanlardan bazıları şunlardır.

- Taşımacılık ve ulaşım
- Turizm ve otelcilik
- Telekominiksayon
- Belediyeler

4. ÇİMENTO SEKTÖRÜNE BİR UYGULAMA

Basma dayanıklılığı en önemli çimento özelliğidir, öyle ki kalite kontrol için ana parametredir [5]. Basma dayanıklılığının belirlenmesi için standart “28 gün basma dayanıklılığı testi” yaygın olarak kullanılır [6]. Bu test çimento üretimi sürecinde her partiden alınan numunelerin 28 gün bekletilerek basma mukavemetin deneysel olarak belirlenmesini içerir. Fakat çimento basma dayanıklılığının deneysel sonuçlarının elde edilmesi için 28 gün beklemek endüstri için uzun bir zamandır. Bu nedenle, basma mukavemetinin daha hızlı belirlenmesi çimento endüstrisi için bir ihtiyaçtır ve araştırmacıların ilgisini hak etmektedir.

Basma mukavemetinin tespiti için esas olarak iki farklı yol vardır: (a) hızlandırılmış dayanıklılık testi yöntemleri ve (b) matematiksel modellerin kullanımı. Bu makalede ikinci yöntem üzerine odaklanılmıştır. Geçmişte en sık kullanılan matematiksel yaklaşım basit regresyon modellerini kullanmaktır [5,6]. Basma mukavemeti kimyasal ve fiziksel çok çeşitli faktörlerle bağlantılıdır. Dayanıklılık üzerinde bu faktörlerin etkisini açıklamak için kullanılan istatistiksel modelleri de içeren analitik modeller çok kompleks olabilirler [7]. Bu nedenle,

basma mukavemetini tahmin etmek için veri madenciliği tekniklerinin kullanılması umut verici bir yaklaşım olarak görülmektedir. Bu çalışmada Portland kompozit çimentosunun basma mukavemetini tahmin etmek için *gen denklem programlama* ve *yapay sinir ağları* olarak bilinen iki yeni nesil veri madenciliği yöntemi ve *regresyon analizi* olarak bilinen klasik bir veri madenciliği yöntemi kullanılarak bu yöntemlerin performansları karşılaştırılmıştır.

Gen denklem programlama, *yapay sinir ağları* ve *regresyon analizi*nde kullanılan veri Adıyaman'da bulunan bir çimento fabrikasından alınmıştır. Toplanan veri 104 günlük bir üretim den örneklenmiştir. Çimento dayanıklılık testi Avrupa Standardı EN 196-1'e göre yapılmıştır. Bu araştırmada kullanılan çimento türü Cem II/B 32.5R Avrupa Standardı EN 197-1'tır. Araştırmada kullanılan verinin biçimi Tablo 1'de açıklanmıştır. *Gen denklem programlama* ve *yapay sinir ağları* algoritmalarına girdi olarak düşünülen 19 değişken bulunmaktadır. Tahmin edilecek sadece bir çıktı (28-gün sonundaki basma mukavemeti) vardır. İlk 79 günlük üretim verisi algoritmaların eğitim verisi olarak kullanılmıştır, 25 günlük üretim verisi ise algoritmalar için test verisi olarak kullanılmıştır. Regresyon analizinde de benzer bir yaklaşım izlenmiştir.

4.1. Gen denklem programlama kullanarak model oluşturma ve analiz

Genetik algoritmalar ailesinin yeni nesil bireylerinden olan gen denklem programlama veriler arasındaki ilişkileri evrimsel mekanizmaları kullanarak öğrenmeye çalışan ve bulabildiği yapıları denklemler şeklinde ifade edebilen bir algoritmadır. Algoritma ağaç türü veri yapılarını listeler halinde gösterebilen çok etkili bir denklem türetme mekanizması kullanılır. Bazı genetik operatörlerle sürekli değişime uğratılan denklemler problem verisi en iyi ifade edebilecek forma doğru optimize edilirler. Gen denklem programlama ile ayrıntılı bilgiler literatürde bulunabilir [8].

Portland kompozit çimentonun basma mukavemetini modellemek için Tablo 2'de anlatılan veri formatı kullanılmıştır. Modellemedeki esas amaç girdi değişkenlerini (d_1, d_2, d_3, …….d_{18}, d_{19}) ve çıktı değişkenini (y) bağlayan gizli fonksiyonu tanımlamaktır. Bu fonksiyon genel olarak $y = f(d_1, d_2, d_3, \ldots\ldots.d_{18}, d_{19})$ formunda gösterilebilir. Gen denklem programlama algoritması ile elde edilen fonksiyon çimentonun 28-gün sonundaki basma dayanımını tahmin etmede kullanılacaktır. Algoritmada kullanılan parametreler Tablo 3'de mevcuttur.

Tablo 2. Model yapımında kullanılan değişkenler

Kod	Girdi değişkeni	Kod	Çıktı değişkeni
d1	SiO_2 (%)	*y*	28-gün mukavemet (MPa)
d2	Loss on ignition (%)		
d3	Al_2O_3 (%)		
d4	Fe_2O_3 (%)		
d5	CaO (%)		
d6	MgO (%)		
d7	SO_3 (%)		
d8	Alumina Modulus		
d9	K_2O (%)		
d10	Serbest lime (%)		
d11	Litre ağırlık(l/g)		
d12	Percentage of Composite (%)		
d13	Sieve residue on 45μm (%)		
d14	Sieve residue on 90μm (%)		
d15	Specific Surface (cm^2/g)		
d16	Ayarlama zamanı (min)		
d17	1 gün mukavemet CCS (MPa)		
d18	2 gün mukavemet CCS (MPa)		
d19	7 gün mukavemet CCS (MPa)		

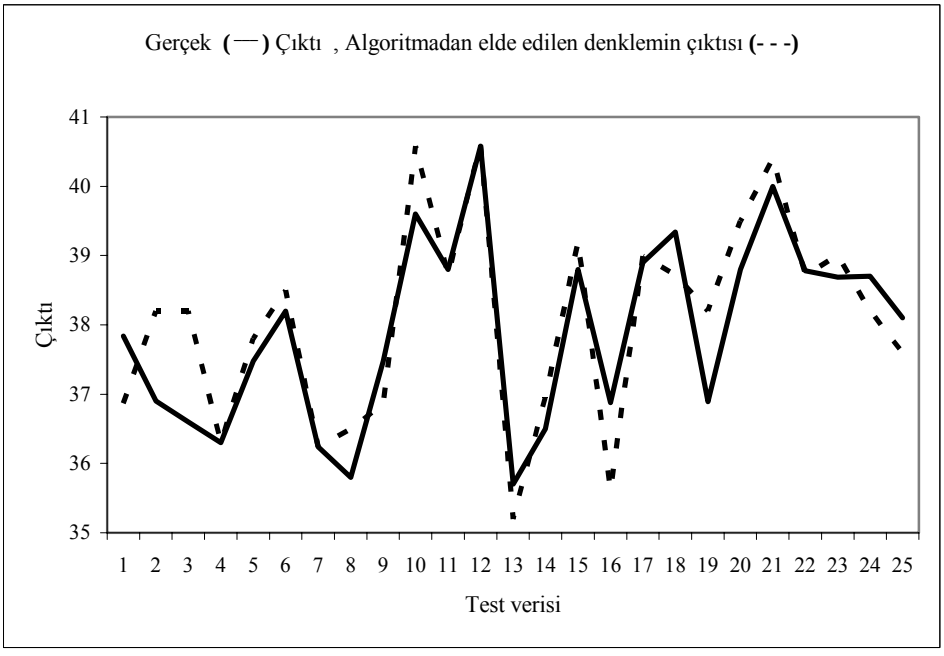
Tablo 3. Gen denklem programlama algoritmasında kullanılan bazı parametreler

p1	İterasyon sayısı	3000 – 20000 arasında
p2	Populasyon büyüklüğü	50
p3	Fonksiyon kümesi	+, −, ∗, /, ^ , √ , e, log, Sin, Tan
p4	Genlerin sayısı	1, 2, 3
p5	Baş büyüklüğü	5, 8, 10
p6	Bağlantı fonksiyonu	+, *
p7	Mutasyon oranı	0.014, 0.044, 0.084

Gen denklem programlama algoritmasından elde edilen en iyi sonuç 0.775 R^2 hatasına sahiptir. GEP algoritması tarafından çimentonun basma mukavemeti tahmininde kullanılmak üzere en elde edilen en iyi denklem eşitlik 1 de gösterilmiştir.

$$y = d_5 - \log(Tan(d_{10}) + d_{18}) + \sqrt{d_{18} * d_4 - Tan(d_{17})} \quad (1)$$

En iyi denklem için elde edilen test sonuçları Şekil 8'de gösterilmiştir. Şekil 8'de de görülebileceği gibi, tahmin edilen ve gerçek mukavemet eğrileri arasında iyi bir paralellik mevcuttur. Algoritma ile elde edilen denklem gerçek verinin eğilimini çok yakından takip edebilmektedir.



Şekil 8. Çimento dayanıklılık tahmini için gen denklem programlama algoritmasının değerlendirmesi

4.2. Yapay sinir ağları kullanarak model oluşturma ve analiz

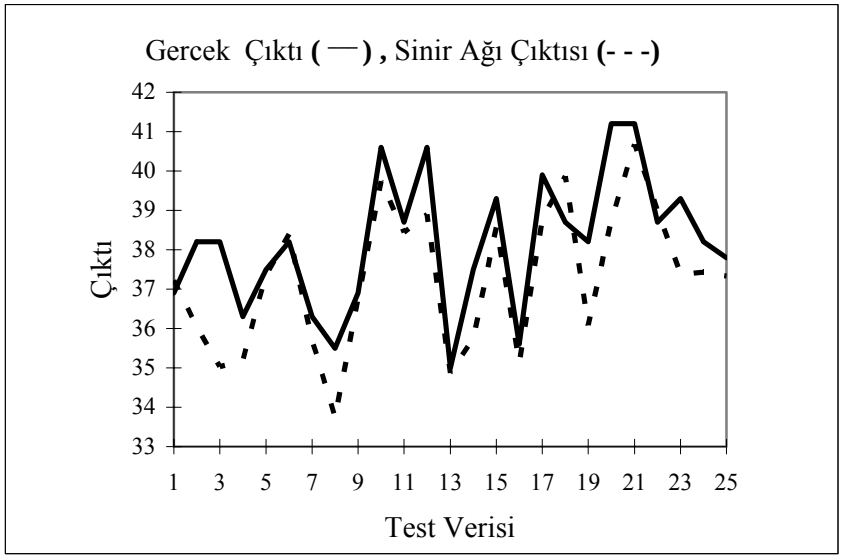
Portland kompozit çimentonun basma mukavemeti tahmini için yedi farklı sinir ağı yapısı kullanılmıştır. Bunlar *multilayer perceptron*, *generalised feedforward*, *modular network*, *Jordan/Elman*, *self-organizing map*, *principal component analysis* ve *recurrent network*. olarak bilinen ağ yapılarıdır. Tablo 2'de tanımlanan deneysel veri yapısı Portland kompozit çimentonun 28-gün basma mukavemetinin modellemesi için kullanılmıştır. Asıl amaç, girdi değişkenleri (d_1, d_2, d_3, …….d_{18}, d_{19}) ve çıktı değişkeni (y) arasındaki ilişkiyi gösterebilen eğitilmiş sinir ağını ve bu ağın ağırlık katsayılarına ulaşmaktır. Sinir ağını eğitmek için Delta-Bar-Delta algoritması [9] kullanılmıştır. Bu algoritma sinir ağlarını eğitmek için kullanılan en iyi eğitim algoritmalarından biridir [9]. Elde edilen eğitilmiş sinir ağı çimentonun 28-gün basma mukavemeti tahmininde kullanılmıştır. Sinir ağı testlerinde kullanılan parametreler Tablo 4'de verilmiştir.

Tablo 4. Sinir ağı algoritmalarının parametreleri

1	Gizli Katmanların Sayısı	1, 2, 3
2	Gizli Katmanlardaki Nöronların Sayısı*	7, 10, 13
3	Sinir Ağlarının Türü	Multilayer Perceptron, Generalised Feedforward, Modular Network, Jordan/Elman, Self-Organizing Map, Principal Component Analysis, Recurrent Network

*Gizli katmanlardaki nöronların sayısının ortalama değeri **2*KareKök(Girdi sayısı + 1) ile tahmin edilmiştir** [10]. Diğer değerler ortalama değere 3 ekleyerek ve çıkararak elde edilmiştir.

Sinir ağı algoritmasından elde edilen en iyi sonuç 0.696 R^2 hataya sahiptir. Bir gizli katmanlı ve gizli katmanda 13 nöron olan sinir ağı en iyi sonucu üretmiştir. Bu sinir ağı için elde edilen test sonucu Şekil 9'da gösterilmiştir. Sinir ağından elde edilen en iyi sonuç gen denklem programlama algoritmasından elde edilen en iyi sonucundan yaklaşık olarak %10 daha kötüdür.



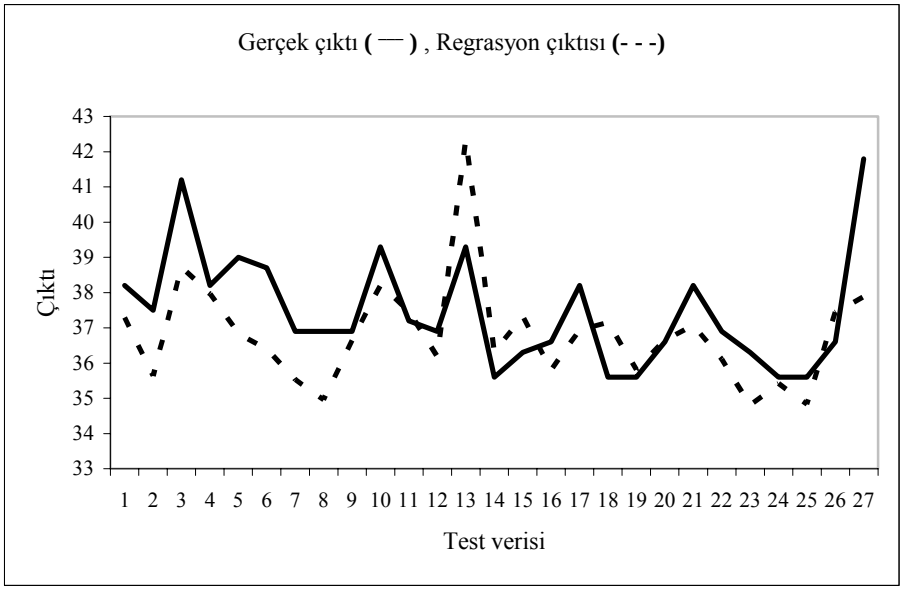
Şekil 9. Çimento dayanıklılık tahmini için sinir ağı metodunun değerlendirmesi

4.3. Regresyon analizi

Klasik istatistiksel yaklaşımlara göre gen denklem programlama ve sinir ağı tekniklerin tahmin gücü hakkında fikir sahibi olmak amacıyla regresyon analizi tekniği uygulanmıştır. Tablo 2'de tanımlanan deneysel veri yapısı regresyon modelinde de kullanılmıştır. Asıl amaç girdi değişkenlerini (d_1, d_2, d_3, ……d_{18}, d_{19}) çıktı değişkenine (y) bağlayan çok-değişkenli regresyon denklemini belirlemektir. Elde edilen denklem Portland kompozit çimentonun 28-gün basma mukavemeti tahmininde kullanılacaktır. Regrasyon analizini gerçekleştirmek için SPSS yazılım paketi kullanılmıştır. Bu analizden elde edilen regresyon denklemi eşitlik 2 de verilmiştir.

$$\begin{aligned} y = {} & 129.2629 - 1.06461d_1 + 0.228794d_2 + 0.197611d_3 - 4.97767d_4 - 0.58978d_5 + 0.349407d_6 \\ & - 2.05767d_7 - 3.03088d_8 - 12.0376d_9 + 0.861342d_{10} + 0.00072d_{11} + 0.191208d_{12} \\ & - 0.19533d_{13} - 0.80748d_{14} - 0.00296d_{15} - 0.03243d_{16} + 0.039985d_{17} - 0.07229d_{18} \\ & + 0.602732d_{19} \end{aligned} \quad (2)$$

Test verisi üzerinde regresyon denklemi ile yapılan tahminin R^2 si 0.357'dir. Test verisi üzerinde regresyon denklemi ile elde edilen maksimum ve minimum hata yüzdeleri %9.38 ve %0.15'dir. Regresyon analizinin test sonuçları Şekil 10'da gösterilmiştir. Sonuçlardan görüldüğü gibi regresyon analizi gen denklem programlama ve sinir ağı algoritmalardan çok daha kötü sonuç vermiştir. Çimentonun basma mukavemeti birçok kimyasal ve fiziksel faktöre dayandığı ve bu faktörler arasındaki ilişkinin kompleks ve doğrusal olmadığı için bu beklenen bir sonuçtur.



Şekil 10. Çimento dayanıklılık tahmini için regresyon analizinin değerlendirmesi

6. SONUÇLAR

Bu makalede veri madenciliği tekniği, yöntemleri ve uygulama alanları tartışılarak çimento sektöründe yapılan bir uygulama çalışması tanıtıldı. Yapılan çalışma çimentonun 28 günlük basma mukavemetinin gen denklem programlama, yapay sinir ağları ve regresyon analizi ile tahminini içermektedir. Yapılan çalışma sonucunda yapay zeka temelli yöntemlerin daha iyi sonuç verdiği gözlenmiştir. Özellikle gen denklem programlama diğer yöntemlerden daha iyi sonuç vermiştir.

REFERANSLAR

[1] Vahaplar, A., İnceoğlu, M.M., Veri madenciliği ve elektronik ticaret, http://www.bayar.edu.tr/bid/dokumanlar/inceoglu.doc , Son erişme 20 Oçak 2004.

[2] Fayyad, U. M., Piatetsky-Shapiro, G., Uthurusamy, R. (1996), Advances in Knowdedge Discovery and Data Mining. Cambridge, MA, MIT Press.

[3] Veri madencilik uygulamaları ve eylimi, http://courses.cs.deu.edu.tr/cse572/Veri%20MadencilikUygulamalar.doc , Son erişme 20 Oçak 2004.

[4] Robert Cooley, Bamshad Mobasher, Jaideep Srivastava, "Web Mining: Information and Pattern Discovery on the World Wide Web", http://www-users.cs.umn.edu/~mobasher/webminer/survey/survey.html, Son erişme 20 Oçak 2004.

[5] Tango, C.E.S., (1998), An extrapolation method for compressive strength prediction of hydraulic cement products, Cement and Concrete Research, 28, 969-983.

[6] Tsivilis, S., Parissakis, G., (1995), A mathematical-model for the prediction of cement strength, Cement and Concrete Research, 25, 9-14.

[7] Akkurt, S., Ozdemir, S., Tayfur, G., Akyol, B., (2003), The use of GA-ANNs in the modelling of compressive strength of cement mortar, Cement and Concrete Research, 33, 973-979.

[8] Ferreira, C., (2001), Gene expression programming: a new adaptive algorithm for solving problems,Complex Systems, 13, 87-129.

[9] Elmas, Ç., (2003), Yapay Sinir Ağları, Seçkin Yayıncılık, Ankara.

[10] NeuNet Pro 2.2, http://www.cormactech.com/neunet, Son erişme 20 Oçak 2004.