KADİR HAS ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

BÖBREK NAKLİ GEÇİRMİŞ HASTALARDA AKILLI YÖNTEM TABANLI YENİ ÖZNİTELİK SEÇME ALGORİTMASI GELİŞTİRİLMESİ

Çağıl ACAR ŞAYLAN

Ocak, 2013

ÇAĞIL ACAR ŞAYLAN

M.S. TEZ

2013

2013

KADİR HAS ÜNİVERSİTESİ FEN BİLİMLERİ ENSTİTÜSÜ

BÖBREK NAKLİ GEÇİRMİŞ HASTALARDA AKILLI YÖNTEM TABANLI YENİ ÖZNİTELİK SEÇME ALGORİTMSI GELİŞTİRİLMESİ

ÇAĞIL ACAR ŞAYLAN

JÜRİ ÜYELERİ:

Prof.Dr. Hasan DAĞ (Tez Danışmanı) : _____

Yrd.Doç.Dr. Songül ALBAYRAK : _____

Yrd.Doç.Dr. Öznur YAŞAR DİNER : _____

SUNUM TARİHİ: 17.01.2013

TEŞEKKÜR

Çalışmalarım boyunca değerli yardım ve katkılarıyla beni yönlendiren tez danışmanım Prof. Dr. Hasan DAĞ'a, yine kıymetli tecrübelerinden faydalandığım uygulama çalışması için yol gösteren ve her aşamada yardımcı olan hocam Işıl YENİDOĞAN'a teşekkür ederim.

Tez yazımı sürecinde yardımlarını esirgemeyen Dr. Oğuzhan CEYLAN'a teşekkür ederim.

Attığım her adımda manevi destekleriyle, sevgi ve sabırlarıyla beni hiçbir zaman yalnız bırakmayan çok değerli aileme ve eşime teşekkürü bir borç bilirim.

Çağıl Acar Şaylan

Ocak 2013

İÇİNDEKİLER

Sayfa

KISALTMALAR

ARFF **:** Dosya Formatı (**A**ttribute **R**elation **F**ile **F**ormat)
BT **: B**ilgisayarlı **T**omografi
UCI **: U**niversity of **C**alifornia , **I**rvine
GA **:** Genetik Algoritması (**G**enetic **S**earch)
HS **:** Harmoni Arama (**H**armony **S**earch)
KAK **: K**oroner **A**rter **K**alsifikasyonu
P **:** Fosfor
PSO **:**Parçacık Sürüsü Optimizasyonu (**P**artical **S**warm **O**ptimization)
SVM **:** Destek Vektör Makinası **(S**upport **V**ector **M**achine**)**
ToT **:** Nakilden bu yana geçen zaman (**T**ime **o**n **T**ransplantation)

ÇİZELGE LİSTESİ

Sayfa

ŞEKİL LİSTESİ

Sayfa

BÖBREK NAKLİ GEÇİRMİŞ HASTALARDA AKILLI YÖNTEM TABANLI YENİ ÖZNİTELİK SEÇME ALGORİTMASI GELİŞTİRİLMESİ

ÖZET

Veri madenciliği, verilerden keşfedilecek desenler yardımıyla yeni bilgiler elde etme amacıyla çok farklı disiplinlerde kullanılan çeşitli metotlardan oluşmaktadır. Tıp alanındaki verinin büyüklüğü ve hayati önem taşıması, Veri madenciliğinin bu alanda da uygulanmasını gerekli kılmıştır. Bu tezde Veri Madenciliğinin Tıp alanında kullanımı incelenmiştir. Uygulama çalışması için İstanbul Üniversitesi Cerrahpaşa Tıp Fakültesi'nde ayakta tedavi gören hastalar arasından, Mart 2006 – Aralık 2007 tarihleri arasında 21 aylık bir sürede tedavisi görmüş hastalara ait veriler bir araya getirilerek bir veri kümesi oluşturulmuştur. Bu veri kümesi üzerinde WEKA yazılımı kullanılarak sınıflama, kümeleme ve karar ağacı algoritmaları çalıştırılmış, elde edilen karar kuralları uzman desteğiyle incelenerek koroner arterlerde kalsifikasyon bulunmasında etkili olan faktörlerin neler olduğu belirlenmiş ve öznitelik seçme algoritmalarıyla aynı faktörlere ulaşılıp ulaşılamadığı belirlenmiştir**.**

Naive Bayesian sınıflandırma algoritması kullanılarak, çapraz geçerlilik ölçütü (cross validation) amaç fonksiyonu olarak alınan armoni araması (harmony search) yöntemiyle eniyileme problemi çözülmüştür. Yazılan program Weka da hazırlanmış, hasta verilerini alınıp Matlab girdi veri olarak kullanılmıştır. Eniyileme, sınıflandırma ve çapraz geçerlilik ölçütü programları Matlab yazılımı kullanılarak yazılmıştır.

INTELLIGENT METHOD BASED ON NEW FEATURE SELECTION ALGORITHM ON RENAL TRANSPLANTATION PATIENTS

SUMMARY

Data mining consists up of many different methods which try to find new information from data patterns. This is the main reason why it has been a basis for many research areas. The amount of data which belongs to the field of medicine is extensive and also very significant this is the main reason behind why the usage of data mining on these types of datasets have been needed. In this thesis the usage of data mining on the field of medicine has been investigated. The dataset consists of the data from the outpatients of the University of Istanbul - Cerrahpasa Medical Faculty which were treated throughout the period of 21 months between the dates March 2006 - December 2007. With the aid of the WEKA software the dataset was examined with classification, clustering and decision tree algorithms and some decision rules where found. These decision rules were then analysed with the help of specialists to determine which features caused complications in the coronary arteries. Also a comparison with feature selection algorithms were done to see if the same features could be found.

We solve optimization problem by using Harmony Search algorithm, by taking cross validation results as objective function and using Naive Bayesian classification algorithm. Our program gets the patient data prepared in Weka and uses it as input to Matlab, a commercial package developed for performing calculations using matrix operations. Optimization, classification and cross validation modules were programmed in Matlab.

1.GİRİŞ

Bilgisayar ve iletişim teknolojilerindeki gelişmelere paralel olarak donanımın ucuzlaması verilerin uzun süre depolanmasına dolayısıyla büyük kapasiteli veritabanlarının oluşmasına neden olmuştur [1]. Teknolojinin büyük bir hızla gelişmesi, durmadan büyüyen ve işlenmediği sürece atıl kalan veri yığınlarını meydana getirmiştir. Değersiz gibi görünen bu veri yığınlarında anlamlı ilişkiler kurmak ve değerli bilgiye ulaşabilmek için kullanılan yöntem, temelinde istatistik yatan veri madenciliğidir.

Veri madenciliği araç ve teknolojileri, belirli bir alanda ve özellikle belli bir amaç için toplanmış büyük veri setleri içinde anlamlı, ilginç fakat gizli örüntü ve/veya ilişkileri bulmayı sağlar. Bu amaca ulaşabilmek için kullanıcı merkezli bir yaklaşımla hem istatistiksel modellemeyi, matematiksel algoritmaları ve yapay sinir ağları, karar ağaçları gibi performansı eğitime dayalı otomatik öğrenme metodolojilerini hem de analiz ve tahmin yöntemlerini kullanır.” Veri madenciliği tekniklerinin beraberinde getirdiği en büyük kısıtlama, sanılanın aksine teknolojik değil kullanıcı merkezli yaklaşımın kaçınılmaz sonucu olarak üzerinde çalışılan veri setinin özelliklerine hakim bir uzmanın görüş ve önerilerine mutlak ihtiyaç duyuyor olmasıdır”[2]. “Tıp alanında veri madenciliği teknikleriyle anlamlı ilişkiler kurabilmek ve/veya ilginç, gizli örüntü ve ilişkilere ulaşabilmek ancak tıp konusunda bilgi ve tecrübeye sahip uzmanların yorumlarıyla mümkün olabilir” [3].

1.1 Tezin Amacı

Bu çalışmada veri madenciliğinin tıp alanında kullanımı incelenmiş ve Cerrahpaşa Tıp Fakültesi Hematoloji ABD’ndan elde edilmiş özgün veri seti üzerinde öznitelik seçme algoritmalarının sonuçları karşılaştırılarak en iyileme yöntem tabanlı yeni bir öznitelik seçme algoritması oluşturulması amaçlanmıştır.

Uygulama çalışması kapsamında, böbrek nakli geçirmiş 178 hastaya ait veri setine veri madenciliği teknikleri uygulanarak bu tür hastalar arasında koroner arter kalsifikasyon bulunmasında etkin olan faktörlerin WEKA programı yardımıyla neler olduğu belirlendikten sonra en iyileme yöntem tabanlı öznitelik seçme algoritması

geliştirilerek aynı sonuçların elde edilip edilmediği kontrol edilmiştir. Bu amaçla hazırlanan tez aşağıdaki bölümlerden oluşmuştur :

İkinci bölümde, veri madenciliği kavramı, teknikleri ve tıp alanında kullanımı incelenmiştir.

Üçüncü bölümde veri madenciliğinde daha verimli sonuç elde etmek için kullanılan ayrıklaştırma yöntemini açıklanmıştır.

Dördüncü bölümde veri madenciliğinde spesifik bir konu olan öznitelik seçme algoritmaları ve bu algoritmaların birbirleriyle olan ilişkilerinden bahsedilmiştir.

Beşinci ve altıncı bölümlerde yeni öznitelik seçme algoritması geliştirme çalışmaları ve akıllı yöntem tabanlı sistemler incelenmiştir.

Son bölümde ise, elde edilen sonuçlar irdelenerek katkılar vurgulanmıştır. Ayrıca ileride yapılabilecek benzeri çalışmalar ve uygulama alanları için öneriler tartışılmıştır.

.

2.VERİ MADENCİLİĞİ

2.1 Veri Madenciliğinin Tarihçesi

1950'li yıllarda matematikçiler veri madenciliği teknikleri üzerine çalışarak mantık ve bilgisayar bilimleri alanlarında yapay zeka ve makina öğrenme alanlarını ortaya çıkarmışlardır. 1960'lı yıllarda istatistikçiler yeni bir algoritma keşfetmişlerdir. Veri madenciliğinin ilk adımlarını oluşturan bu algoritmalar regresyon analizi ve en büyük olabilirlik kestirimidir. Daha sonraki 20 yıllık süreçte önce verilerin sınıflara ayrılması ardından bu sınıflar arasında ilişkisel bağlantıların kurulması ile veri tabanı kavramı ortaya çıkarılmıştır. 1990'lı yıllara gelindiğinde ise veritabanında bilgi keşifinin ilk adımları oluşturulmuş ve bunula birlikte büyük veritabanları için veri ambarı geliştirilmiştir. Aynı zamanlarda yeni teknolojilerle beraber veri madenciliği yaygın olarak kullanılmaya başlanmıştır.

Shalvi ve DeClaris'e (1998) göre " Veri madenciliği, belirli bir alanda ve belirli bir amaç için toplanan veriler arasındaki gizli kalmış ilişkilerin (desenlerin, modellerin vb.) ortaya konulmasıdır" [4].

Veri madenciliğinin detaylı bir tanımı şöyledir:
"Veri madenciliği, depolanmış yüksek miktardaki veriden istatistiksel ve matematiksel teknikler gibi desen tanımlayıcı teknolojiler kullanarak anlamlı ve yeni ilişkiler, desenler ve trendler keşfetme prosesidir" [5].

"Veriler içinde desen aramak insan hayatının başlamasından beri süre gelen bir süreçtir. Bilim insanlarının işi ise veriler içinde fiziksel dünyanın nasıl çalıştığını anlatan modeller aramak ve bulacağı modeller yardımıyla ortaya çıkacak yeni durumlarda neler olacağını tahmin edecek teoriler geliştirmektir" [6].

İnsanoğlu geçmişten bugüne kadar veri madenciliği tekniklerini geliştirmek için her zaman verileri yorumlayıp bilgi edinmeye çalışmıştır ve bunun için çeşitli donanımlar oluşturmaya başlamışlardır. Şekil 2.1'de kronolojik olarak veri madenciliğinin gelişim süreci verilmiştir.

Gelişim Adımları	Cevaplanan Karar Problemi	Kullanılabilen Teknolojiler	Ürün Sağlayıcıları	Karakteristikler
Veri Toplama (1960'lar)	"Benim toplam karım geçen 5 yılda ne kadardı?"	Bilgisayarlar, Teypler, Diskler	IBM,CDC	Geriye dönük , statik veri dağıtımı
Veri Erişimi (1980'ler)	"İngiltere'de geçen mart ayında birim satışları ne kadardı?"	İlişkisel Veritabanları, SQL, ODBC	Oracle,Sybase, Informix,IBM, Microsoft	Kayıt düzeyinde geriye dönük, dinamik veri dağıtımı
Veri Ambarlama ve Karar Destek Sistemleri (1990'lar)	"İngiltere'de geçen mart ayında birim satışları ne kadardı?"	OLAP, Çok Boyutlu Veritabanı Sistemleri, Veri ambarları	Pilot, Comshare, Arbor,Cognos, Microstrategy	Çoklu düzeylerde, geriye dönük dinamik veri dağıtımı
Veri Madenciliği (Bugün)	"Gelecek ay Boston'daki birim satışlar muhtemelen ne olabilir, niçin?	İleri düzeyde algoritmalar, çok işlemcili bilgisayarlar, büyük veritabanları	Pilot, Lockheed, IBM,SGI, SPSS,SAS, Microsoft vs.	Geleceğe dönük ,proaktif enformasyon dağıtımı

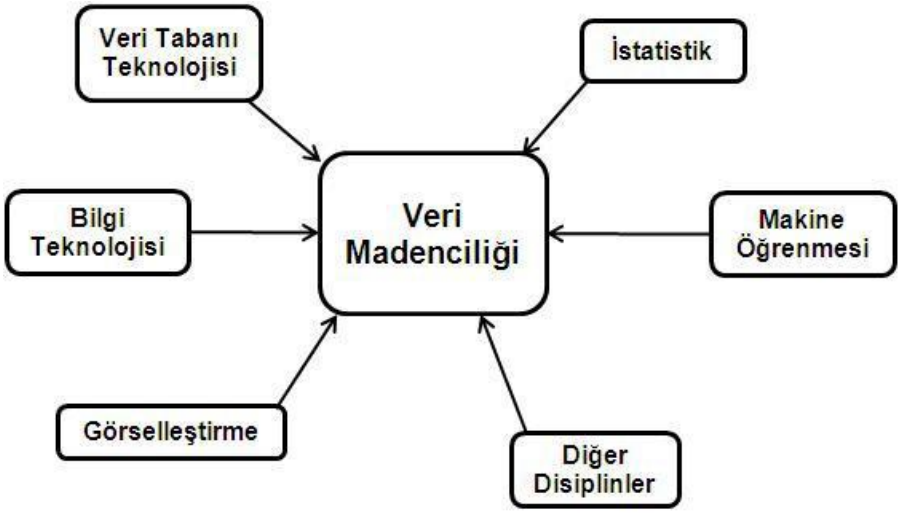
Şekil 2.1 : Veri Madenciliği gelişim süreci

2.2 Veri Madenciliği ile İlişkili Bilim Dalları

Veri madenciliği farklı disiplerinlerden faydalanan disiplinlerarası bir alandır. Veri madenciliği; veri tabanı sistemleri,istatistik, makine öğretimi, insan-makine etkileşimi, veri görselleştirme gibi farklı disiplinler kümesinden oluşmaktadır.

Veri madenciliği yaklaşımına bağlı olarak, veri madenciliği'nde farklı disiplinlerde kullanılan tekniklerden de yararlanılmaktadır. Yararlanılan tekniklere; sinirsel ağlar, bulanık küme teorisi gibi örnekler verilebilir.

İşlenen verinin yapısına bağlı olarak veri madenciliği, bilgisayar grafikleri, uzaysal veri analizi, web teknolojileri, resim analizi, gibi tekniklerden de faydalanmaktadır.



Şekil 2.2 : Veri Madenciliği ile diğer disiplinler arası ilişki [7].

2.3 Veri Madenciliğinin Kullanıldığı Sahalar

"Veri madenciliği verinin yoğun olarak üretildiği hemen her ortamda uygulama alanı bulmuştur. Veri madenciliğinin uygulama alanlarını şu şekilde sıralayabiliriz.

Satış ve Pazarlama alanında veri madenciliği uygulamarı,

a) Müşteri sınıflandırma, hedef müşteri belirleme

b)Müşterilerin demografik özellikleri arasındaki bağlantıların bulunması

c) Pazar sepeti analizi

d)Müşteri ilişkileri yönetimi

e) Satış tahmini yapmak amacıyla kullanılır.

Bankacılık alanında veri madenciliği uygulamaları,

a)Farklı finansal göstergeler arasında gizli korelasyonların bulunması

b) Kredi kartı harcamalarına göre müşteri gruplarının belirlenmesi

c) Müşteri kredi risk araştırmaları

d) Kredi kartı dolandırıcılıklarının tespiti

e) Hisse senetlerinin değer değişimi ile ilgili tahminler yapılması için kullanılır

Sigortacılık alanında ise veri madenciliği uygulamaları şu alanda kullanılmaktadır,

a) Poliçe onaylama

Bu uygulamaların dışında taşımacılık ve ulaşım, konaklama ve kamu alanlarında da veri madenciliği uygulamarı yapılmaktadır ” [8].

2.4 Veri Madenciliğinin Faydaları

Veri madenciliği rekabetin fazla olduğu piyasalarda, firmaların konumlarını sağlamlaştırmak adına birtakım değerlerinin yönetilmesinde etkili rol oynamaktadır. Veri Madenciliğinin faydalarını şu şekilde sıralayabiliriz:

i. Müşterilerin elde tutulmasına yardımcı olur.

ii. Müşteri davranışlarının anlaşılmasını sağlar.

iii. Sigortacılık, bankacılık ve telekomünikasyon alanlarında geçmiş veriler kullanılarak sahtekarlık yapanlar için bir model oluşturma ve benzer davranışlar gösterenleri belirleme konusunda veri madenciliği önemli bir rol oynar.

2.5 Veri Madenciliği Yazılımları

Veri madenciliği metotlarının uygulanmasını kolaylaştırmak amacıyla bu metotları uygulayabilen çeşitli bilgisayar yazılımları geliştirilmiştir.

“Kurumsal amaçlarla kullanılmak üzere uygun bir veri madenciliği yazılımı seçmek için dikkat edilmesi gereken hususlar aşağıda verilmiştir :

- Yazılımın uygulayabildiği veri madenciliği teknikleri
- Yazılımın kapasite kısıtları (Veri boyutu, kullanıcı sayısı, veri kümesi içinde özellik sayısı, uygun donanım tipi)
- Veri tabanlarına ve dosyalara erişim imkanı
- Çok boyutlu kullanıcı arayüzü desteği ve kullanım kolaylığı
- Oluşturulan modellere ilişkin detaylı açıklama desteği
- Görsel sunum, raporlama formatlarının yeterliliği
- Farklı yazılımlar ile etkileşim uyumu
- Danışma ve eğitim desteği

Bazı ticari veri madenciliği yazılımlarının güçlü ve zayıf nitelikleri Çizelge 2.1’de özetlenmiştir” [9].

Çizelge 2.1: Veri Madenciliği Yazılımlarının güçlü ve zayıf nitelikleri[10]

ÜRÜN	GÜÇLÜ NİTELİĞİ	ZAYIF NİTELİĞİ
Clementine	Görsel Arayüzü, Algoritma genişliği	Ölçeklenirlik
Darwin	Etkili hizmet sağlayıcı, sezgisel arayüz opsiyonu	Sınırlı görselleştirme
DataCruncher	Kullanım kolaylığı	Tek algoritma içermesi
Enterprise Miner	Algoritma derinliği, görsel arayüzü	Kullanımı zor, kompleks yapısı
GainSmarts	Veri dönüştürme, algoritma seçeneği genişliği	Sınırlı görselleştirme
Intelligent Miner	Algoritma genişliği, grafiksel ağaç/küme çıktısı	Algoritma seçeneği az
MineSet	Veri Görselleştirme	Algoritma seçeneği az, model ihracına izin vermez
Model1	Kullanım kolaylığı, otomatik model keşfi	Dikey bir araç olması
ModelQuest	Algoritma genişliği	Sezgisel olmayan arayüz seçenekleri
PRW	Kapsamlı algoritmalar, otomatik model seçimi	Sınırlı görselleştirme
CART	Ağaç opsiyonları derinliği	Dosya açılması zorluğu, sınırlı görselleştirme
Scenario	Kullanım kolaylığı	Dar analiz yolu
NeuroShell	Çoklu Sinirsel ağ mimarisi	Alışılmışın dışında arayüz, sadece sinirsel ağ içermesi
OLPARS	Çoklu istatistiksel algoritmalar, sınıf tabanlı görselleştirme	Eski arayüz, Dosya açılması zorluğu
See-5	Ağaç opsiyonları derinliği	Sınırlı görselleştirme, az veri seçeneği
S-Plus	Algoritma derinliği, görselleştirme, programlanabilir	Sınırlı çıkarım metodu, dik öğrenme eğrisi
WizWhy	Kullanım ve model kavrama kolaylığı	Sınırlı görselleştirme

Tez kapsamında yapılan uygulama çalışmasında Weka version 3.6 veri madenciliği yazılımı kullanılmıştır.

2.5.1 Veri Madenciliği Yazılım Paketi : WEKA

"Waikato Üniversitesi tarafından "Weka Makine Öğrenmesi Projesi" için hazırlanmış olup, bünyesinde birçok özeğitimli makine algoritmasını barındıran açık kaynak kodlu genel kamu lisansına sahip Java tabanlı bir veri madenciliği yazılım paketidir" [11]. Verilerin daha düzenli ve daha sıralı olmasını sağlayan bu format aslında veritabanlarını metin tabanında taklit etmektedir.

WEKA çalıştırıldıktan sonra Şekil 2.3'de görüldüğü gibi, Application menüsünde çalışılabilecek modlar listelenmektedir. Bunlar komut modunda çalışmayı sağlayan

Simple CLI, projeyi adım adım görsel ortamda gerçekleştirmeyi sağlayan Explorer ve projeyi sürükle bırak yöntemiyle gerçekleştirmeyi sağlayan KnowledgeFlow seçenekleridir.

Şekil 2.3 : WEKA'da Applications Menüsü

WEKA'ya import edilen üç farklı dosya formatı bulunmaktadır bunlar; Arff, Csv ve C4.5'dur. WEKA'nın içerisinde veri işleme, veri sınıflandırma, veri kümeleme, veri ilişkilendirme özellikleri mevcuttur. Bu adımdan sonra yapılacak olan projenin amacına göre açılan sayfadaki uygun tabdaki (Sınıflandırma, Kümeleme, İlişkilendirme) uygun algoritma veya algoritmalar seçilerek veriler üzerine uygulanmakta ve en doğru sonucu veren algoritma seçilebilmektedir.

2.5.2 ARFF

ARFF makine öğrenmesinde kullanılan bir formattır. ARFF aracılığıyla okunan veriler programlama seviyesinde karakter dizileri biçiminde temsil edilir. Dosya yapısını belirlemek için @relation, @attribute ve @data deyimleri kullanılır. @relation deyimi veri yığının genel amacını ya da ismini belirtir. @attribute deyimi ise verideki veri tabanında sütunlara karşılık gelen özellik isimlerini belirtmek için kullanılırken, @data deyimi ham verilerin başladığı satıra işaret eder.

2.6 Veri Madenciliğinin Tıptaki Önemi ve Uygulamarı

Veri madenciliği telekomünikasyon ve pazarlama gibi alanlarda başarısını kanıtladıktan sonra insanların sağlık sorunlarını çözmeye çalışan tıp alanında da uygulanmaya başlanmıştır.

Tıp alanında veri madenciliği tekniklerinin uygulanması için gerekli altyapı, veri tabanı uygulamalarının hastane ve laboratuarlarda kullanılmaya başlanması ile sağlanmıştır.

Böylece bilgisayarlar hasta bakım hizmetlerinin destekleme, sağlık bakım hizmetlerinin kalitesinin değerlendirilmesi gibi doğrudan sağlık bakım hizmetlerinin sunulmasında kullanılmasının yanı sıra, karar verme, yönetim,
planlama ve tıbbi araştırmalar gibi yönetsel ve akademik fonksiyonların yerine getirilmesinde daha fazla kullanılmaya başlanmıştır.

Basit veri tabanları hasta bilgileri (ad, soyad, adres, kan grubu vb.) gibi temel verileri saklarken, günümüzde hastların hasta yerleşiminden finansal verilerine, ilaçlarından ziyaret planına kadar pek çok veri de bu sistemlerce kayıt altına alınmaktadır [12].

2.6.1 Literatür Özeti

Tıp alanında Veri Madenciliği çalışmaları gün geçtikçe hızla çoğalmaktadır.

Tıp alanında veri madenciliği tıbbı teşhislerde ve uygun tedavi sürecinin belirlenmesi gibi birçok alanda kullanılmaktadır. Veri madenciliği teknikleri günümüzde pazarlama sektöründen, bankacılık ve sigortacılık gibi alanlarda elektronik ticaret ile ilgili alanlarda ve tıbbi alanlarda yaygın bir şekilde kullanılmaktadır. Tıbbi alanda; kalp-damar hastalıklarından [13], kanser hastalıklarına (gögüs, prostat vb), diyabet, troid vb alanlara uygulanmış, hatta hastane yönetimine ve tedavi yöntemlerinin optimize edilmesine kadar çeşitlendirilmiştir.

Tıbbi veri setleri üzerinde yapılan başka bir çalışmada; Cheng ve arkadaşları tarafından yapılmış olan tıbbi verilerde veri madenciliği için öznitelik seçmelerinin otomatik yaklaşımlarının ve uzman kararlarının karşılaştırılmasıdır. Kardiyovasküler

hasta veri setinde iki öznitelik seçme yaklaşımıyla risk aralığının sınıflama başarısını değerlendirmişlerdir [14].

Günümüzde özellikle karmaşıklık derecesi yüksek olan problemlerde (Np hard problemlerde) sayısal enyileme yöntemleri bazı durumlarda yetersiz kalmaktadır. Bu yöntemlere alternatif olarak akıllı yöntemler gündeme getirilmiştir. Bu yöntemlere örnek olarak benzetilmiş tavlama algoritması (ing: simulated annealing) [15], genetik algoritmalar (ing: genetic algorithms) [16], parçacık sürüsü eniyilemesi (ing: particle swarm optimization)[17], armoni araması (ing: harmony search) [18] gibi yöntemler verilebilir. Söz konusu yöntemlerde bir ilk aday çözüm kümesi oluşturulur ve doğa esinli yaklaşımlar kullanılarak bu ilk aday çözümler daha iyi çözümlere taşınır. Sözkonusu yöntemlerin en yenilerinden bir tanesi de caz müzisyenlerinin doğaçlama yaparken daha iyi notalar bulmasının benzetimini yaparak eniyileme problemini çözen armoni araması yöntemidir [18].

Bir başka çalışmada Srinivas ve ekibi, kalp hastalıklarının modern toplumlarda en sık rastlanılan ve en fazla ölüme sebebiyet veren hastalık oluşundan yola çıkıp spesifik olarak Andhra Pradesh (Hindistan)'te Singareni isimli kömür madenciliği bölgesinde bildirilen kardiyovasküler hastalık oranının, diğer risk faktörleri de gözönüne alındıktan sonra diğer bölgelerle mukayesesi sonucu gerçekten fazla olup olmadığını incelemişlerdir.

Bağımlı değişkenler olarak kalp-damarla ilgili teşhisleri veya kardiyovasküler hastalıklara işaret eden

1)Göğüs ağrısı
2)İnme ve
3)Kalp Krizi'ni belirlemişlerdir.

Kalple ilgili araştırmalar hastalığın varlığının tahmini ile ilgili 15 farklı özellik belirlemiştir. Düzenli kriterler dışında daha genel kriterler olan Beden Kütle İndeksi, hekim görüşü, yaş, etnik köken, eğitim seviyesi, gelir düzeyi ve buna benzer diğer kriterler de tahmin yürütmede kullanılmıştır. Tıbbi teşhislerde bir karar destek mekanizmasının oluşturulması kaliteyi arttırırken maliyetleri de azaltacaktır. Bu noktadan hareketle yapılan çalışmada kalp hastalıklarıyla ilgili isabetli tahmin

yürütme için popüler veri madenciliği tekniklerinden Karar Ağaçları, Naive-Bayes ve Yapay Sinir Ağları algoritmaları kullanılmıştır [19].

2.6.2 Tıp Alanında Veri Madenciliği Uygulamarı

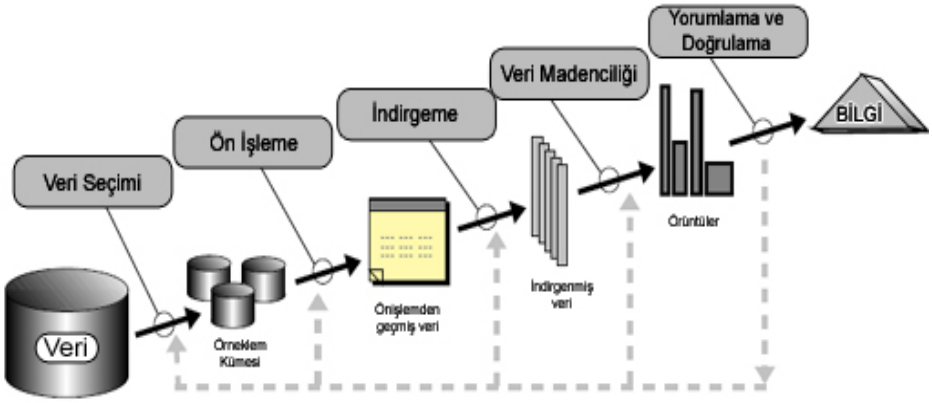
Tıp alanında veri madenciliği uygulamalarını altı başlık altında anlatabiliriz :

- Tıp Literatürü üzerinde metin madenciliği uygulamaları
- Hastane Bilgi Sistemi üzerinde veri madenciliği çalışmaları
- Genetik alanında veri madenciliği çalışmaları
- Eczacılık alanında veri madenciliği çalışmaları
- Hastalık Çeşitlerine göre yapılan veri madenciliği çalışmaları
- Böbrek Nakli Geçirmiş Hastalarda Koroner Arter Kalsifikasyonunun İncelenmesi çalışmaları

2.7 Veri Madenciliği Süreci

Veri madenciliği teknikleriyle verilerden bilgi keşfi süreci aşağıdaki adımlardan oluşmaktadır:

1. Verilerin Toplanması
2. Verilerin Temizlenmesi
3. Verilerin Bütünleştirilmesi
4. Verilerin Dönüştürülmesi
5. Veri Madenciliği
6. Desen Değerlendirme
7. Bilgi Sunumu [7]



Şekil 2.4 : Veri Madenciliği Aşamları

Veri madenciliğini daha detaylı bir şekilde ele almak istersek aşağıdaki adımaları izlememiz gerekmektedir.

i)İşin (Problemin) Analizi ve Verilerin Anlaşılması

ii) Veri seçimi

iii) Veri analizi ve hazırlığı

iv) Veri azaltma ve dönüştürme

v) Önemli özelliklerin (Attributes) seçimi

vi) Veri kümesi boyutunun küçültülmesi

vii) Normalleştirme

viii) Birleştirme

ix) Veri madenciliği metodunun seçimi

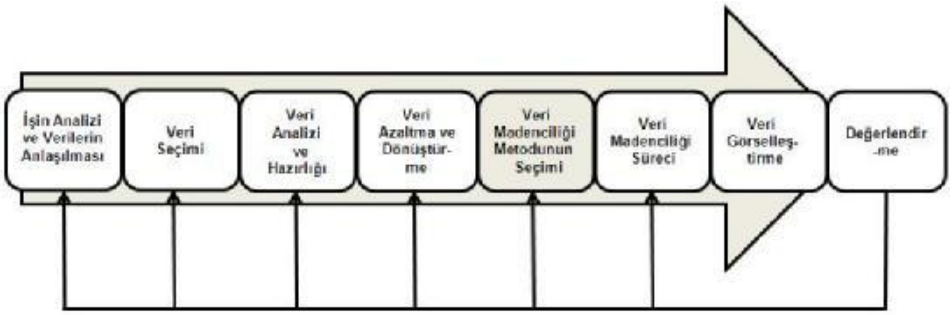
x) Veri madenciliği süreci

xi) Görselleştirme

xii) Değerlendirme

xiii) Bilgi kullanımı ve sonuçların hedefe uygun olarak değerlendirilmesi

Şekil 2.5'de Veri madenciliği sürecinin detaylı adımları gösterilmiştir.



Şekil 2.5 : Veri Madenciliği süreci

Veri madenciliğinin dar tanımı ve geniş anlamı dikkate alınarak oluşturulan genel deneysel prosedür Şekil 2.6'da belirtilen 5 adımdan oluşmaktadır.

Şekil 2.6 : Veri Madenciliği süreci adımları [20]

2.7.1 Verilerin Toplanması

Bu aşama verilerin nasıl toplanacağı ile ilglidir. Verilerin oluşturulması yani toplanması sürecinde iki farklı yaklaşım vardır. Eğer süreç uzman kontrolünde yapılırsa tasarlanmış deney; uzman kontrolü olmadan yapılırsa gözlemsel yaklaşım olarak adlandırılır.

Kullanılan verilerin aynı bilinmeyen örneklemden gelmesi, modelin kurulması, test edilmesi ve uygulanması açısından önemlidir.

2.7.2 Problemin Belirlenmesi ve Verilerin Anlaşılması

Problemin belirlenmesi ve verilerin anlaşılması kısımı veri madenciliği uygulmasının ilk aşamasını oluşturmaktadır. "Problemin belirlenmesi aşamasında bilinmeyen bağımlılıklara göre değişkenler belirlenir ve bir model oluşturmak için veriler arası ilişkilerden hipotez veya hipotezler oluşturulmaya çalışılır"[20]. Veri madenciliği uygulamalarında problemin iyi belirlenmesi gerekmektedir. Problemin belirlenmesi kullanılacak tekniğin seçilmesinden daha önemlidir. Problemin

belirlenmesi ve verilerin anlaşılması adımının en iyi şekilde gerçekleştirilebilmesi için aşağıda belirtilen kurallara uyulması sayesinde ulaşılabilir:

i) Problem elle tutulur faydalar sağlayacak şekilde açıkça tanımlanmalı.
ii) Muhtemel sonuç belirlenmeli
iii) Elde edilecek sonucun nasıl kullanılacağı belirlenmeli.
iv) Problem ve veriler olabildiğince anlaşılmalı.
v) Problem modele dönüştürülmeli
vi) Varsayımlar belirlenmeli.
vii) Model döngüsel olarak iyileştrilmeli.
viii) Model mümkün olan en basit hale getirilmeli.
ix) Modelin kararsızlığı tanımlanmalı.
x) Modelin belirsizliği tanımlanmalı.

Bu belirtilen kurallara uyulduktan sonra problemin ve verilerin niteliğine bağlı olarak ilgili alanda uzman desteğine ihtiyaç duyulabilir ve uzman ile veri madenciliği çalışmalarını yürüten kişilerle yapılan iş birliği sayesinde süreç daha başarılı işlenebilir.

2.7.3 Verilerin Hazırlanması

"Veri madenciliği sürecinde zamanın en fazla harcandığı yer olan verilerin hazırlanması kısmı sürecin başarılı olmasında %75 ila %90 arasinda bir katkı sağlar. Veri kümesinin zayıf veya var olmayan verilerden hazırlanması sürecin başarısızlığında %100 sorumludur [16]".

Verilerin hazırlanma süreci, veri madenciliği sürecinin diğer adımlarından bağımsız değildir. "Veri madenciliğinin her adımında yapılan tüm işlemler birlikte yeni ve daha gelişmiş bir veri kümesi elde etmemize yardımcı olur [20]". Verilerin hazırlanması aşaması; gerçek hayattan toplanan eksik, geniş

dağılımlı, çelişkili verilerin temizlenmensi, verilerin kaynaştırılması ve dönüşümü, veri hacminin küçütülmesi, verilerin kesikli hale getirlmesi işlemlerinden oluşur.

Verilerin temizlenmesi

Verilerin temizlenmesi aşaması, eksik değerlerin doldurulması, uyumsuz verilerin tanımlanması, çelişkili verilerin doğrulanması veya veri setinden çıkartılması gibi işlemlerden oluşur. "Bu aşamada eksik verilerin yerine yenileri belirlenerek konulması için aşağıda belirtilen yöntemler kullanılabilir :
a)Eksik değer içeren kayıtlar veri kümesinden atılabilir.
b)Kayıp değerlerin yerine bir genel sabit kullanılabilir.
c)Değişkenin tüm verileri kullanarak ortalaması hesaplanır ve eksik değer yerine bu değer kullanılabilir.
d)Değişkenin tüm verileri yerine, sadece bir sınıfa ait örneklerin değişken ortalaması hesaplanarak eksik değer yerine kullanılabilir"[21].

Veri setinde bulunan uyumsuz veriler (Outliers) gözlemlere uymayan anormal veri değeri olarak adlandırılır. Bu tarz veriler genelde ölçüm ve yazım hatalarından kaynaklanır. Uyumsuz verilerle başa çıkmak için verilerin temizlenmesi adımının bir parçası olarak uyumsuz verilerin tespiti gerekmektedir.

Verilerin kaynaştırılması ve dönüştürülmesi

Verilerin kaynaştırılması aşamasında farklı kaynaklardan elde edilen verilerin uyumlu bir forma sokulması için bazı işlemlerin yapılması gerekmektedir. Bu işlemeleri şu metadata, korelasyon analizi, veri çatışması tespiti ve semantik uyumsuzluğun giderilmesi gibi aşamalardan oluşur. Verilerin dönüştürülmesi kısmı ise verilerin, veri madenciliği uygulamasında daha iyi sonuç verecek forma sokulması işlemidir.

Veri dönüştürme işleminin sonucunda elde edilen değer, model için bir önem ifade etmelidir. Böylece elde edilen değer ayrı bir özellik olarak tanımlanabilir. " Örnek olarak tıbbi veri kümelerinde sıklıkla rastlanan hasta ağırlık ve boy değerleri kullanılarak veri dönüştürme yoluyla Vücut Kitle İndeksi denen bir orana çevrilir. Bu oran yardımıyla hastaların kilolu olup olmadığının tespiti daha iyi yapılabilmektedir" [20].

Verilerin kesikli hale getirilmesi ve kavramsal hiyerarşi oluşturma

Verilerin kesikli hale getirilmesi sürekli değişkenler üzerinde uygulanan bir durumdur.Verilerin kesikli hale getirilmesi için şu teknikler kullanılmaktadır; birleştirme, histogram analizi ve entropi 'dir. Uygulanacak algoritmanın tipine bağlı olarak verilerin dönüştürülmesi işlemi gerçekleştirilir.

Verilerin dönüştürülmesi aşamasında sayısal tabanlı algoritma kullanılacak ise kesikli verilerin sürekli hale getirilmesi; kategorik tabanlı algoritmalar kullanılacaksa sürekli verilerin kesikli hale getirilmesi gerekmektedir.

2.7.4 Modelin Kurulması

Tanımlanan problem için uygun modelin belirlenmesi aşamasında, mümkün olduğunca çok sayıda model kurularak ve bu kurulan modellerin denenmesi ile gerçekleşir. Bu nedenle veri hazırlama ve model kurma aşamaları, en iyi olduğu düşünülen modele ulaşıncaya kadar yinelenen bir aşamadır.

2.7.5 Modelin Yorumlanması

Model elde edilmesi beklenen hedefleri karşılamaya yeterli bulunduğu zaman, süreç bazlı daha geniş bir perspektiften değerlendirme yapılır. Bu değerlendirme sürecinde modelin doğru kurulup kurulmadığı, gelecekte kullanılabilecek farklı verilerin neler olabileceği, modelin genişletilmesi gibi konuları içerir.

3. AYRIKLAŞTIRMA

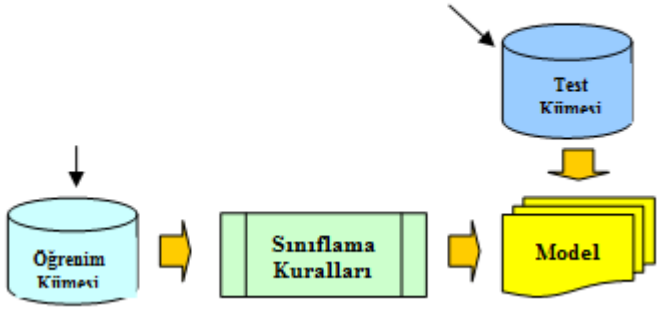
3.1 Ayrıklaştırma Nedir

Ayrıklaştırma için model kuruluş süreci, denetimli ve denetimsiz öğrenmenin kullanıldığı modellere göre farklılık göstermektedir.

Ayrıklaştırma sürecini Akpınar şu şekilde açıklamıştır "Örnekten öğrenme olarak da isimlendirilen denetimli öğrenmede, bir denetçi tarafından ilgili sınıflar önceden belirlenen bir kritere göre ayrılarak, her sınıf için çeşitli örnekler verilir. Sistemin amacı verilen örneklerden hareket ederek her bir sınıfa ilişkin özelliklerin bulunması ve bu özelliklerin kural cümleleri ile ifade edilmesidir.

Öğrenme süreci tamamlandığında, tanımlanan kural cümleleri verilen yeni örneklere uygulanır ve yeni örneklerin hangi sınıfa ait olduğu kurulan model tarafından belirlenir.

Denetimsiz öğrenmede, kümeleme analizinde olduğu gibi ilgili örneklerin gözlenmesi ve bu örneklerin özellikleri arasındaki benzerliklerden hareket ederek sınıfların tanımlanması amaçlanmaktadır".



Şekil 3.1 : Denetimli Öğrenme

“Denetimli öğrenmede seçilen algoritmaya uygun olarak ilgili veriler hazırlandıktan sonra, ilk aşamada verinin bir kısmı modelin öğrenilmesi, diğer kısmı ise modelin geçerliliğinin test edilmesi için ayrılır. Modelin öğrenilmesi, öğrenim kümesi kullanılarak gerçekleştirildikten sonra test kümesi ile modelin doğruluk derecesi belirlenir”.

“Bir modelin doğruluğunun test edilmesinde kullanılan en basit yöntem basit geçerlilik testidir. Bu yöntemde tipik olarak verilerin % 5 ile % 33 arasındaki bir kısmı test verileri olarak ayrılır ve kalan kısım üzerinde modelin öğrenimi gerçekleştirildikten sonra, bu veriler üzerinde test işlemi yapılır. Bir sınıflama modelinde yanlış olarak sınıflanan olay sayısının, tüm olay sayısına bölünmesi ile hata oranı, doğru olarak sınıflanan olay sayısının tüm olay sayısına bölünmesi ile ise doğruluk oranı hesaplanır.”
(Doğruluk Oranı = 1 - Hata Oranı)

Sınırlı miktarda veriye sahip olunması durumunda, kullanılabilecek diğer bir yöntem, çapraz geçerlilik testidir. Bu yöntemde veri kümesi rastgele iki eşit parçaya ayrılır. İlk aşamada bir parça üzerinde model eğitimi ve diğer parça üzerinde test işlemi; ikinci aşamada ise ikinci parça üzerinde model eğitimi ve birinci parça üzerinde test işlemi yapılarak elde edilen hata oranlarının ortalaması kullanılır. Bir kaç bin veya daha az satırdan meydana gelen küçük veri tabanlarında, verilerin n gruba ayrıldığı n katlı çapraz geçerlilik testi tercih edilebilir. Verilerin örneğin 10 gruba ayrıldığı bu yöntemde, ilk aşamada birinci grup test, diğer gruplar öğrenim için kullanılır. Bu süreç her defasında bir grubun test, diğer grupların öğrenim amaçlı kullanılması ile sürdürülür. Elde edilen on hata oranının ortalaması kurulan modelin tahmini hata oranı olacaktır.

.

4.ÖZNİTELİK SEÇME ALGORİTMALARI

4.1 Bilgi Kazancı

Bir T elemanını tanımak için gereken bilgi ile X özelliği de kullanılarak T elemanını tanımak için gereken bilgi arasındaki farktır. Bu aynı zamanda, X özelliğine dayalı bilgi kazancı olarak adlandırılır ve aşağıdaki gibi ifade edilir.

Kazanç(X,T)= Bilgi (T)- Bilgi (X, T) (4.1)

Örnek : "Bir özellik için bilgi kazancının hesaplanması
{A1,A2,...,An} değerlerine sahip A özelliği ağacın bölünmesi için kullandığında,T kümesi de {T1,T2,...,Tn} şeklinde bölünecektir. Bu bölümlemede T kümesindeki A özelliğinin Ai olduğu bölgelere Ti kümesi densin. Bu kümedeki pozitif olayların sayısı pi, negatif olayların sayısı ni olarak gösterildiğinde Ti alt ağacı için beklenen bilgi gereksinimi I(pi,ni) olur. T ağacı için beklenen bilgi kazancı tüm Ti ağaçlarının beklenen bilgi kazançlarının ağırlıklı ortalamalarının toplamı olur ve

$$E(A) = \sum_{i-1}^{n} \frac{pi + ni}{p + n} I(pi, ni) \tag{4.2}$$

şeklinde hesaplanır. Dolayısı ile A özelliği üzerinden sağlanan bilgi kazancı aşağıdaki eşitlikle ifade edilir." [22]

Bilgi kazancı (A) = I(p,n) – E(A) (4.3)

4.2 Kazanım Oranı

Bilgi kazanımı metodu çok çeşitli değerlere sahip nitelikleri seçme eğilimindedir. Bu problemi çözmek için C4.5 kazanım oranı kullanılmaktadır.

$$SpliInfo_A(D) = -\sum_{j=1}^{v} \frac{|D_j|}{|D|} x log_2 \left(\frac{|D_j|}{|D|}\right) \quad (4.4)$$

$$GainRatio(A) = Gain(A) / SplitInfo(A) \quad (4.5)$$

Böylelikle en yüksek kazanım oranına sahip nitelik seçilmiş olur.

4.3 Korelasyon Tabanlı Özellik Seçici

"Korelasyon Tabanlı Özellik Seçici (Correlation-based Feature Selection), Weka içerisinde yer alan özellik seçici metotlardan biridir. Diğer özelliklerle düşük korelasyonlu, sınıf değişkeni ile yüksek korelasyonlu olan özellikleri seçer. Eğer iki niteliğin sahip olduğu değerler birbirleri ile simetrik olarak değişmekteyse, bu iki nitelik birbiri ile ilişkilidir. Diğer durumda ise bu iki nitelik için birbiri ile ilişkisiz olduğu kabul edilir." [23]

5. AKILLI YÖNTEM TABANLI YENİ ÖZNİTELİK SEÇME ALGORİTMASININ GELİŞTİRİLMESİNDE KULLANILAN YÖNTEMLER

5.1 Çalışmanın Uygulama Alanı

Bu bölümde akıllı yöntemler olan parçacık sürüsü eniyileme yöntemi, armoni araması yöntemi ve anlatılacaktır.

5.2 J48 Karar Ağacı ve Navie Bayes Sınıflandırma Algortimaları

Bayes sınıflandırıcısı istatistiksel sınıflandırma teknikleri arasıda yer alır. Bayes sınıflandırıcısı basit bir olasılıksal sınıflandırıcıdır ve eldeki verilere göre hipotezlerin doğru olma olasılığına göre hareket eder. Bir sonucun çıkma olasılığı o sonucu etkileyen tüm faktörlerin o sonucu sağlama olasılıklarının çarpımıdır.

J48 karar ağacı ise C4.5 algoritmasının biraz benzeridir. Entropy'i hesaplayarak karar ağaçları çıkarır ve her bir öznitelik için bilgi kazacını hesaplayarak veri setini böler. Karar verebilmek için en fazla kazanç sağlayan öznitelik kullanılır.

5.3 Parçacık Sürüsü Optimizasyon Algoritması

"Parçacık Sürüsü Algoritması (PSO); 1995 yılında J.Kennedy ve R.C.Eberhart tarafından; kuş sürülerinin davranışlarından esinlenilerek geliştirilmiş populasyon tabanlı stokastik optimizasyon tekniğidir [24]." Parçacık Sürüsü Optimizasyon algoritmaları, Genetik Algoritmalar (GA) gibi evrimsel hesaplama teknikleri ile benzerlik gösterir. Parçacık sürüsü optimizasyonu bir probleme çözüm aramaya benzetilen kuş sürülerinin uzayda yerini bilmedikleri yiyeceği aramaları problemidir.

"Kuşlar yiyecek ararken yiyeceğe en yakın olan kuşu takip ederler. Parçacık olarak adlandırılan her tekil çözüm, arama uzayındaki bir kuştur. Parçacık hareket ettiğinde, kendi koordinatlarını bir fonksiyona gönderir ve böylece parçacığın uygunluk değeri ölçülmüş olur. Bir parçacık, koordinatlarını, hızını şimdiye kadar elde ettiği en iyi uygunluk değerini ve bu değeri elde ettiği koordinatları hatırlamalıdır. Çözüm uzayındaki her boyuttaki hızının ve yönünün her seferinde nasıl değişeceği, komşularının en iyi koordinatları ve kendi kişisel en iyi koordinatlarının bir birleşimi olacaktır."[25].

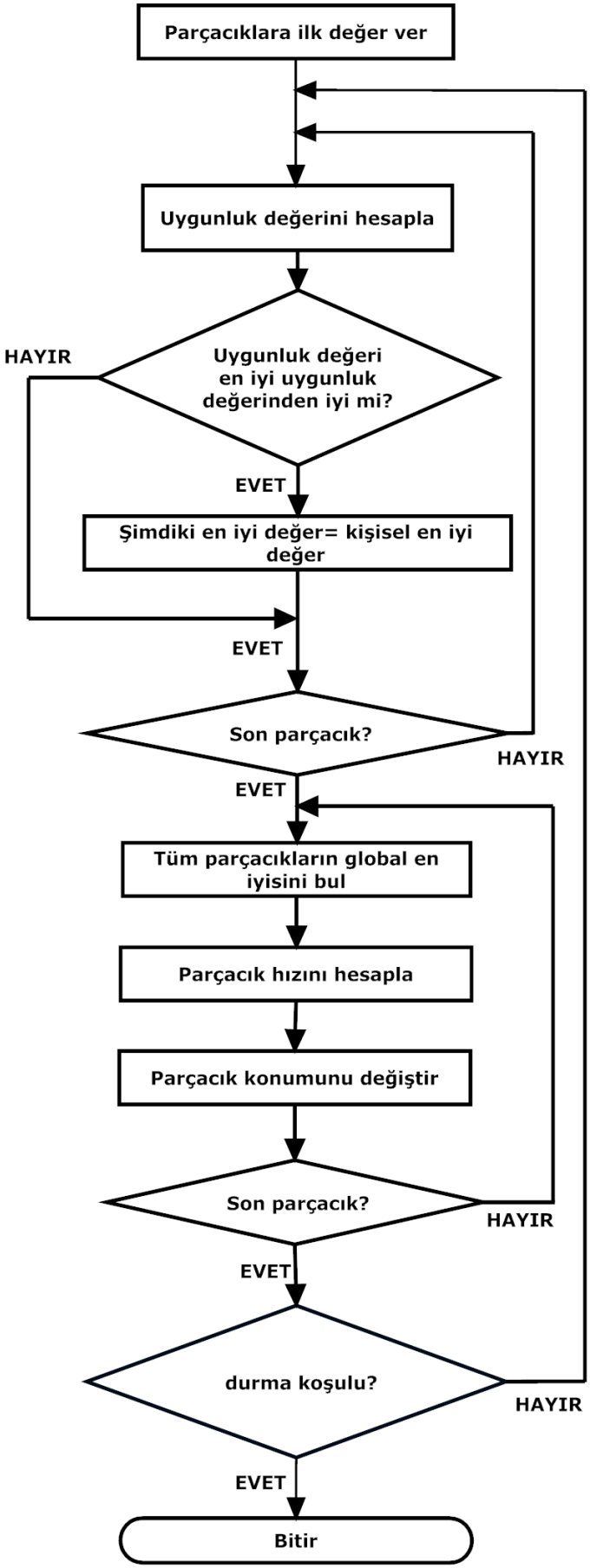
“PSO bir grup rasgele üretilmiş parçacıkla başlar ve iterasyonlar güncellenerek en uygun değer araştırılır. Her iterasyonda, her bir parçacık iki “*en iyi*” değere göre güncellenir. Bunlardan birincisi bir parçacığın o ana kadar bulduğu en iyi uygunluk değeridir. Ayrıca bu değer daha sonra kullanılmak üzere hafıza tutulur ve “*pbest”* yani parçacığın en iyi değeri olarak isimlendirilir. Diğer en iyi değer ise popülâsyondaki herhangi bir parçacık tarafından o ana kadar elde edilmiş en iyi uygunluk değerine sahip çözümdür. Bu değer popülasyon için global en iyi değer olup “*gbest*” olarak isimlendirilir [25]”.

PSO algoritmasının Pseudocode kodu aşağıda verilmiştir :

For *her parçacık için baslangıç kosullamaları*
End
Do For *her parçacık için uygunluk degerini hesapla*
eger uygunluk degeri, pbest ten daha iyi
ise; simdiki degeri yeni pbest olarak ayarla
End

Tüm parçacıkların bulduğu pbest degerlerinin en iyisini, tüm parçacıkların gbest'i olarak ayarla
For *her parçacık için*
(2) denklemine göre parçacık hızını hesapla
(3)denklemine göre parçacık pozisyonunu güncelle
End
While *maksimum iterasyon sayısına veya minimum hata kosulu saglanana kadar devem et*

Parçacık sürüsü eniyileme yönteminin akış şeması Şekil 5.1’de verilmiştir.



Şekil 5.1 : Parçacık sürüsü yöntemi akış şeması.

5.4 Harmoni Arama Algoritması

Harmoni arama algoritması optimizasyon tekniği olarak kullanılan ve ilk olarak Geem ve arkadaşları tarafından geliştirilen HS algoritması, bir orkestradaki müzisyenlerin çaldıkları notalar ile harmonik açıcan en iyi melodinin elde edilmesi prensibine dayanmaktadır[26]. En güzel armoniyi yakalamak, bir anlamda en iyileme problemindeki optimumu bulmaya benzetilmiş ve buna göre yöntem geliştirilmiştir.

Daha iyi bir armoni yakalamaya çalışan müzisyenin önünde üç seçenek vardır. "Bunlardan birincisi kendi hafızasında tuttuğu bir ezgiyi çalmak, ikincisi hafızasındaki bu ezgiyi biraz değiştirerek çalmak ve sonuncusu da yeni ya da rastgele notalar çalmaktır"[27]. Armoni arama yöntemi bu üç seçeneğin bir araya gelmiş halidir.

Armoni arama yönteminin adımları şunlardır :

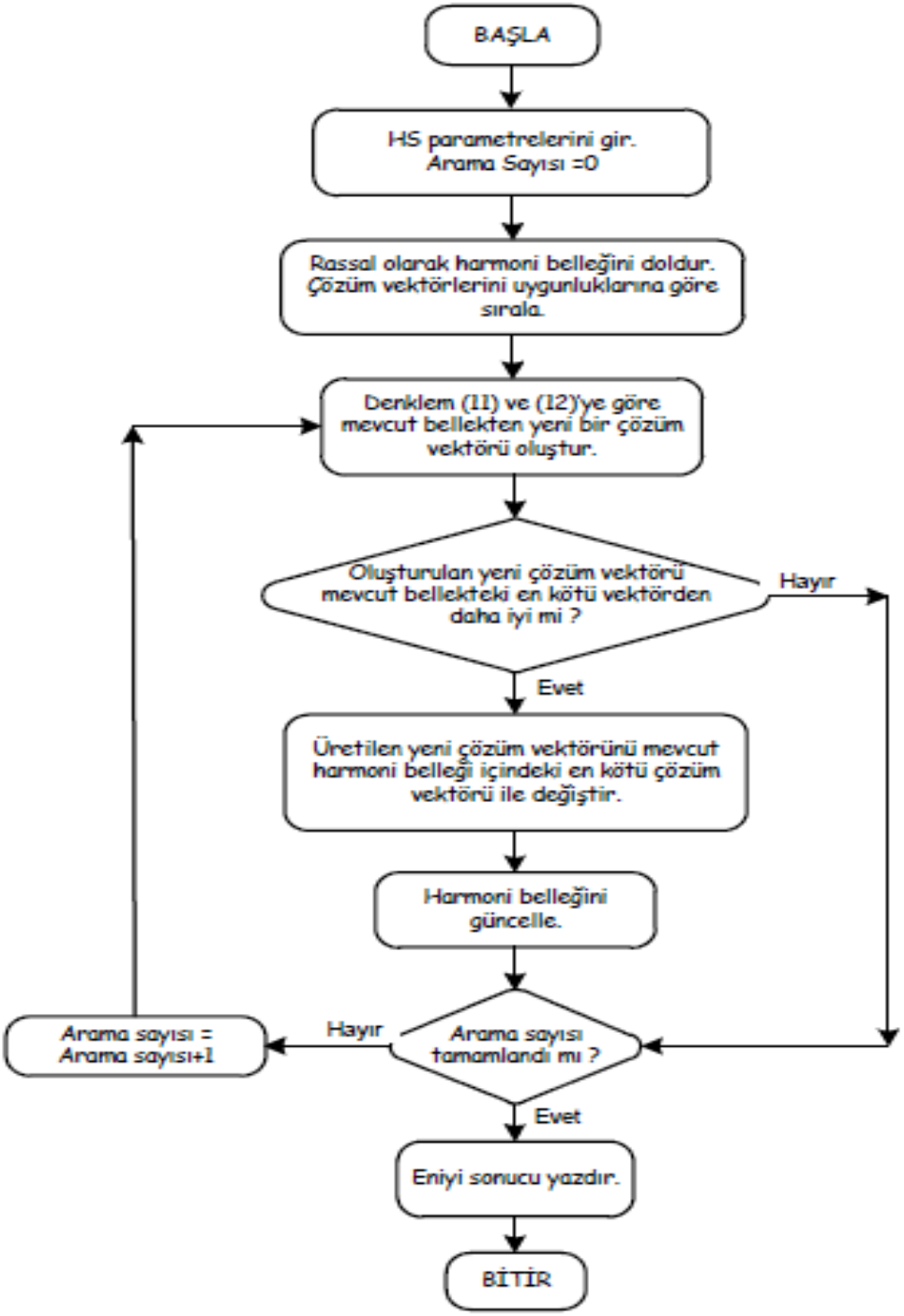
1. Problem ve algoritma parametlerine ilk değerler verilir.
2. Armoni hafızasına ilk değer verilir.
3. Yeni armoni oluşturulur(Doğaçlama).
4. Armoni hafızası yenilenir.
5. Durma koşulunun sağlanıp, sağlanmadığı kontrol edilir.

HS algoritması hesaplama mantığı bakımından genetik algoritmayla benzerlikleri bulunmasına rağmen bu iki yöntem arasındaki en belirgin fark yeniden üretim aşamasındaki varsayımlardan kaynaklanmaktadır.

Genetik algoritmaların kullanımı sırasında toplum içerisindeki iki adet birey kullanılırken, HS metoduyla yeni bir karar değişkeni oluşturulurken toplum içindeki tüm bireylerin özelliklerini taşıyabilmektedir.

HS algoritmasıyla bir optimizasyon problemi beş adımda yapılmaktadır[26]. Bu aşamalar aşağıdaki çizelgede gözterilmiştir.

Çizelge 5.1: Harmoni Arama algoritması şeması



6. AKILLI YÖNTEM TABANLI YENİ ÖZNİTELİK SEÇME ALGORİTMASININ GELİŞTİRİLMESİ

6.1 Veri Setinin Tanıtılması

Veri seti; 8'i kategorik, 18'i sayısal olan toplam 26 öznitelik ve 1 sınıf bilgisinden oluşmaktadır. Sınıf bilgisi ileriki yıllarda "koroner arterlerde kalsifikasyon var (KAK_Var)" ve "koroner arterlerde kalsifikasyon yok (KAK_Yok)" olarak iki değer almaktadır. 178 hasta üzerinde çoklu dedektörlü spiral bilgisayarlı tomografi (BT) ile KAK tespiti yapılmış ve KAK $\geq$ 1 olan hastaları KAK_Var sınıf bilgisiyle etiketlenmiştir. Veri setinde yer alan 178 hastanın %40,4'ünde koroner arterlerde kalsifikasyon bulunurken, %59,6'sında kalsifikasyona rastlanmamıştır [28].

Veri setinde yer alan böbrek nakli geçirmiş 178 hastaya ait özniteliklerin detayları Çizelge 6.1'de verilmiş olup tabloda bulunan sayısal değerler için ortalama değer±standart sapma da gösterilmiştir.

Çizelge 6.1: Böbrek Nakli Geçirmiş Hastalara ait Öznitelikler

Öznit. No	Öznitelik Adı	Açıklama
1	Yaş	Hastanın yaş bilgisi {36,5 ± 11,2}
2	Cinsiyet	Hastanın cinsiyeti {Kadın, Erkek}
3	Nakil Süresi (ay)	Hastanın nakilden sonra BT ile KAK skoru ölçülene kadar geçen süre 70,6±59,5
4	Verici Tipi	Vericinin tipi {Canlı, Kadaverik}
5	Diyaliz Süresi (ay)	Hastanın nakilden önce diyalizle tedavi süresi {24,5 ± 23,5}
6	hs_CRP (mg/L)[a]	Kandaki C Reaktif Protein miktarının hassas ölçümle belirlenmesi {3,1 ± 3,7}
7	Rose_anjina	Hekim kontrolünde yapılan Rose Angina anketi {Evet, Hayır}
8	Sigara Kullanımı	Hastanın sigara içme durumu {Evet, Hayır}

9	Sigara kullanım periyodu (paket/yıl)	Sigara kullanma sıklığı (3 yıldır günde 2 paket sigara içen hasta için 6 paket/yıl) {4,8 ± 9,6}
10	Geçmiş kalp damar hastalıkları	Geçmişte kalp damar hastalıklarından (MI, by-pass, PTCA, CVA) birini geçirip geçirmediği bilgisi {Evet, Hayır}
11	Aile_hikayesi	Birinci derece akrabalarda kalp damar hastalıkları olup olmadığı bilgisi {Evet, Hayır}
12	Vücut kütle endeksi (kg/m^2)	Vücut kütle endeksi {25,7 ± 4,3}
13	Küçük Tansiyon (mmHg)	Diyastolik {79,9 ± 11,1}
14	Büyük Tansiyon (mmHg)	Sistolik {122,6 ± 16,3}
15	Hipertansiyon	Sistolik 140 mm Hg den, Diyastolik 90 mmHg den büyük veya hasta düzenli tansiyon ilacı kullanıyorsa "var" aksi durumda "yok" bilgisi {Var, Yok}
16	T_kol (mg/dL)	Total kolesterol (LDL+HDL+VLDL) {188,8 ± 41,6}
17	LDL-Kolestrol (mg/dL)	Kötü huylu kolestrol {111,8 ± 33,2}
18	HDL-Kolestrol (mg/dL)	İyi huylu kolestrol {49,1 ± 12,3}
19	Trigliserid (mg/dL)	Trigliserid {151,6 ± 76,1}
20	Albuminuri (mg/gün)	24 saatlik idrarda ölçülen albumin miktarı {250,9 ± 586,7}
21	Ca (mg/dL)	Kandaki Kalsiyum konsantrasyonu {9,6 ± 0,5}
22	P (mg/dL)	Kandaki Fosfor konsantrasyonu {3,4 ± 0,7}
23	Ca_P_product (mg^2/dL^2)	Kandaki Kalsiyum Fosfat konsantrasyonun çarpımı (CaxP) {32,2 ± 6,3}
24	PTH (pg/dL)	Kandaki Paratiroid hormonu konsantrasyonu {114,6 ± 113,6}
25	Diabetus mellitus	Diabet hastalığı bilgisi {Evet, Hayır}
26	MDRD (mL/min/1.73 m^2)	Hastanın kreatin ölçümü ve yaşına göre böbrek fonksiyonunu gösteren

		parametre {61,0 ± 20,5}

En iyileme tabanlı öznitelik seçme algoritması geliştirilme süreci aşağıdaki adımlardan oluşmaktadır:

1.Verilerin Temizlenmesi
2.Veri İndirgeme
3.Veri Dönüştürme
4.Öznitelik Seçme Algoritmalarının Uygulanması
5.Uygulanan Öznitelik Seçme Algoritmalarının Performanslarının Karşılaştırılması
6.Sınıflama Algoritmalarının Denenmesi
7.Yeni Öznitelik Seçme Algoritmasının Geliştirilmesi
8.Bilinen Yöntemlerle Yeni Algoritmanın Sınıflayıcı Başarıları Açısından Karşılaştırılması

6.2 Veri Setinin Hazırlanması

Veri seti incelendiğinde, bulunan bazı değerlerin MATLAB'de işlenebilmesi amacıyla aşağıdaki şekilde tekrar düzenlenmiştir.

Cinsiyet: Erkek = 0 , Kadın = 1

Donor_Type : Type_1 = 0, Type_2 = 1

Veri madenciliği teknikleri uygulamaya başlamadan önce veriyle ilgili yapılması gereken hazırlıkların başında kayıp değerler konusu gelmektedir. İlgilenilen veri setinde kayıp değerlerle karşılaşılması durumunda uygulabilecek yaklaşımlar;

1-Bir öznitelik için kayıp değerlerin toplam kayıt sayısına oranı > %30 ise o öznitelikten vazgeçmek

2-Bir öznitelik için kayıp değerlerin toplam kayıt sayısına oranı < %30 ise o öznitelik için kayıp değerleryerine kayıdın ait olduğu sınıf ortalamasını yerleştirmek

Çalışmamıza konu olan CAC verisetinde bulunan **hs_CRP** sütununda 178 kayıtın %5i kayıp değerken, albumin sütununda da 178 kayıtın %12si kayıp değer olarak edinilmiştir. Bu durum 2.çözüm seçeneğine uygun olduğu için kayıp değerler yerine ait olduğu sınıf bilgisine bağlı kalınarak, koroner arterlerinde kalsifikasyon

bulunanların (KAK_Var) ortalaması olan 2.984 ve koroner arterlerinde kalsifikasyon bulunmayanların (KAK_Yok) ortalaması olan 3.068 değerleri girilmiştir.

Albümin verisinde kayıp değerlere ait olduğu sınıf bilgisine bağlı kalınarak, koroner arterlerinden kalsifikasyon bulunanların (KAK_Var) ortalaması 253.924 ve koroner arterlerinde kalsifikasyon bulunmayanların (KAK_Yok) ortalaması olan 250.904 değerleri girilmiştir.

6.3 Korelasyona Dayalı Öznitelik Seçme Algoritması, Bilgi Kazanç ve Kazanç Oranı Öznitelik Seçme Algoritmaları

Hekimlercede bilinirki bir hastayla ilgili hastaya ait birden fazla öznitelik bulunmaktadır. Bunlar hastaların demografik bilgileri, biokimya takip değerleri, ilaç kullanıp kullanmadıkları gibi veri setinin özelliğine bağlı olarak özniteliklerimizi arttırabiliriz.

Bu özniteliklerin hepsi bir hastalığa etki eden değerler midir sorusuna yanıt bulmak için veri madenciliğinde kullanılan öznitelik seçme algoritmalarını veri setimizde kullandık. Sınıf bilgisine erişmede mevcut 26 öznitelikten hangilerinin daha fazla katkı verdiğini bulmamızda yardımcı olan korelasyona dayalı öznitelik seçme algoritması (CFS), bilgi kazanç (InfoGain) ve kazanç oranı (GainRatio) algoritmalarını kullandık. Bu öznitelik seçme yöntemleriyle 26 öznitelikten sınıf bilgisine ulaşmada en fazla katkı veren 5'li , 6'lı ve 7'li alt kümeler oluşturarak belirlenmiş özniteliklerle aynı sınıf bilgisine ulaşıp ulaşamadığımız kontrol edilmiştir.Öznitelik seçme algoritmalarının belirlediği alt kümeleri karşılaştırabilmek için 4 farklı tıbbi veri setinden yararlanılmıştır. Bu tıbbi veri setlerinin bilgileri 6.2 numaralı çizelgede özetlenmiştir

Çizelge 6.2: Veri Setlerinin Özellikleri

Veri Seti	Öznitelik Sayısı	Kayıt Sayısı	Sınıf Bilgisi	
KAK	26	178	KAK Var (72)	KAK Yok (106)
Kalp	13	303	Num <50 (165)	Num >50 (138)

Hepatit	19	155	Ölü (32)	Yaşayan (123)
Hipotroid	29	3772	N (3481) CH (1394)	PH (95), SH (2)

Bu şeklide veri setimizi tekrar düzenleyip 26 özniteliğe sahip olan veri setimiz WEKA programı yardımıyla öznitelik seçme algoritmaları kullanılarak çalıştırıldığında 5 özniteliğe sahip alt küme aşağıdaki gibi seçilmiştir.

- Age
- Time on Transplantation
- Diabetes Mellitus
- Rose Angina
- Donor Type
- Fosfor

Bu çalışmalar sırasında bilgi kazanç (InfoGain) ve kazanç oranı (GainRatio) öznitelik seçme algoritmalarınınaslında birbirleriyle benzer veya ilgili olduklarını gördük. Çizelge 6.3'de veri setlerine göre belirlemiş olduğumuz öznitelikler gösterilmektedir.

Çizelge 6.3: Algoritmalara göre öznitelikler

Veri	BİLGİ KAZANÇ	KAZANÇ ORANI	KORELASYONA DAYALI
KAK	Age	Age	Age
	Time On Transplantation	Diabetes Mellitus	Time On Transplantation
	Diabetes Mellitus	Time On Transplantation	Diabetes Mellitus
	Rose Angina	P	Rose Angina
	P	Rose Angina	Donor Type
	Donor Type	Past Cardiac Disease	P
	Past Cardiac Disease	Donor Type	Hypertension
Kalp	CP	CP	Thal
	Thal	Thal	CA
	CA	CA	Exang
	Oldpeak	Oldpeak	CP
	Exang	Exang	Oldpeak
	Thalach	Thalach	Thalach
	Slope	Slope	Slope
Hepatiti	Albumin	Ascites	Ascites
	Bilirubin	Bilirubin	Albimun
	Ascites	Albimun	Bilirubin

	Spiders Fatigue Histology Malaise	Varices Spiders Fatigue Histology	Spiders Protime Varices Histology
Hipotroid	TSH FTI IT4 T3 TSH measured On Thyroxine Referral Source	TSH FTI IT4 T3 TSH measured On Thyroxine Pregnant	TSH FTI IT4 T3 On Thyroxine Query Hypothyroid Goitre

Bu tabloya göre 3 öznitelik seçme algoritmalarının %80 oranla uyuştukları görülmektedir.

Bu hesaplamalar sonucunda elde edilen verilerle WEKA'da J48 algoritması tekrardan denenip başarı oranları çapraz geçerlilik ölçütü (cross-validation) %62.92 elde edilmiş ve full traning %92.69'dur.

Sınıflama başarılarını J48 karar ağacı ile kontrol ettiğimizde ise gözle görülür bir farkın olmadığı gözlemlenmiştir. Bu sonuçta çizelge 6.4'de sunulmaktadır.

Çizelge 6.4: J48 ile Sınıflama Başarısı

Veri Seti	**Öznitelik Sayısı**		
	5	6	7
KAK	65/65/65	67/66/67	67/68/67
Kalp	80/78/80	78/78/78	78/78/77
Hepatiti	83/83/83	83/83/82	83/83/82
Hipotroid	98/98/99	99/99/99	99/99/99

6 özniteliğe sahip veri setimizde J48 algoritması denendiği zaman çapraz geçerlilik ölçütü cross-validation başarısı %62.35 ve full traning başarısı ise %92.1348'dir.

Bu testler denendikten sonra ayrıklaştırma yöntemi kullanılarak sınıflama yöntemleri olan J48, destekçi vektör makinası (SVM) ve navie bayes algoritmalarıyla ayrıklaştırmanın KAK veri seti üzerindeki etkisi karşılaştırılmıştır.

Çizlge 6.5: Ayrıklaştırmanın etkisinin J48 algoritmasıyla uygulanması

Öznitelik Seçme Algoritmaları	Tür	Öznitelik Sayısı		
		5	6	7
Bilgi Kazanç	Normal	*65*	*67*	*68*
	Ayrıklaştırılmış	*71*	*72*	*71*
Kazanç Oranı	Normal	*65*	*66*	*68*
	Ayrıklaştırılmış	*72*	*72*	*71*
Korelasyona Dayalı	Normal	*65*	*67*	*67*
	Ayrıklaştırılmış	*69*	*70*	*71*

Desktekçi vektör makinası (SVM) sınıflandırmada (Classification) kullanılan etkili ve basit yöntemlerden birisidir. Düzlemde bulunan iki grup arasında bir sınır çizilerek iki grubu ayırmak yani sınıflandırmak mümkündür. Bu sınır iki grubun da üyelerine en uzak olan yerden çizilerek olmalıdır. Bu sınırın nasıl çizileceğinide SVM belirler.

Destekçi vektör makinası kullanılarak KAK veri setinde ayrıklaştırmanın etkisi 6.6 numaralı çizelgede gözterilmiştir.

Çizelge 6.6: Ayrıklaştıranın SVM kullanılarak KAK veri setine etkisi

Öznitelik Seçme Algoritmaları	Tür	Öznitelik Sayısı		
		5	6	7
Bilgi Kazanç	Normal	*69*	*70*	*69*
	Ayrıklaştırılmış	*69*	*68*	*70*
Kazanç Oranı	Normal	*70*	*70*	*70*
	Ayrıklaştırılmış	*69*	*70*	*69*
Korelasyona Dayalı	Normal	*70*	*70*	*69*
	Ayrıklaştırılmış	*69*	*70*	*71*

Navie bayes yöntemi ile ayrıklaştırmanın etkisini kontrol ettiğimizde ise aşağıdaki sonuçlar elde edilmiştir:

Çizelge 6.7: Ayrıklaştırmanın Navie Bayes kullanılarak KAK veri setine etkisi

Öznitelik Seçme Algoritmaları	**Tür**	**Öznitelik Sayısı**		
		5	**6**	**7**
Bilgi Kazanç	Normal	*67*	*68*	*67*
	Ayrıklaştırılmış	*74*	*74*	*74*
Kazanç Oranı	Normal	*67*	*67*	*68*
	Ayrıklaştırılmış	*74*	*74*	*73*
Korelasyona Dayalı	Normal	*67*	*68*	*68*
	Ayrıklaştırılmış	*74*	*74*	*74*

6.4 Harmoni Araması

Öznitelik seçme yöntemleri kullanılarak elde edilen sonuçlardan sonra başarım oranımızı daha da arttırmak için en iyileme yöntem tabanlı öznitelik seçme algoritmasının yazılmasına karar verilmiştir.

Bu çalışmada öznitelik seçme eniyileme problem olarak ortaya konmuştur ve söz konusu problem akıllı yöntemlerden biri olan harmoni araması kullanılarak çözülmeye çalışılmıştır.

Harmoni arama algoritması kullanılarak yazılan eniyileme yöntemiyle problem çözümşeme aşamasında Navie Bayes sınıflayıcısı kullanılmıştır ve çapraz geçerlilik ölçütü (cross validation) amaç fonksiyonu olarak ele alınmıştır.

Harmoni aramasıyla hazırlanan programı MATLAB'de uyguladığımızda hemen hemen aynı sonuçları hatta çalışmayı bir adım daha ileri götürdüğü görülmüştür.

Harmoni aramasıyla KAK veri seti için hazırlanan programın matlab kodu aşağıdaki gibidir .

```
HMCR=0.9; HMS=30; %DE PARAMETRELERİ
PAR=0.8;
subsetsize=7 // istenilen alt küme sayısı olarak değiştirilebilir.
deneme123=loadARFF('26_vy.arff');
[mdata,featureNames,targetNDX,stringVals,relationName]
=weka2matlab(deneme123,{});
hh=1
for i=1:HMS
   deneme=randperm(26);
   populasyon(i,1:subsetsize)=deneme(1:subsetsize);
   populasyon(i,subsetsize+1)=27;
   %populasyon=round(rand(30,5)*25)+1
end

for i=1:HMS
   a='M';
   b=num2str(i);
   ankara=strcat(a,b);
   for m=1:(subsetsize+1)
      mynames{m}=featureNames{populasyon(i,m)};
   end
   opop=mdata(:,populasyon(i,1:(subsetsize+1)));
   yy=matlab2weka(ankara,mynames,opop);
   wekaClassifier = trainWekaClassifier(yy,'trees.J48',{'-D'});
   % wekaClassifier = trainWekaClassifier(yy,'bayes.NaiveBayes',{'-D'});
   predicted=wekaClassify(yy,wekaClassifier);
   actual = yy.attributeToDoubleArray(subsetsize); %java indexes from 0
   jjj=[predicted actual abs(predicted-actual)];
   h=length(find(jjj(:,3)==1));
   populasyon(i,(subsetsize+2))=100*(h/178);
end

itmax=2000 //  tekrarlama adımını istenilen şekilde değiştirebiliriz.
for iteration=1:itmax
   myrandomnumber=rand;
   if myrandomnumber<HMCR
      for i=1:subsetsize%%%%ONEMMM
         indexnew(i)=round(rand*(HMS-1))+1;
         new1(i)=populasyon(indexnew(i),i);
      end
   else
      for i=1:subsetsize
         new1(i)=round(25*(rand))+1;
      end
```

```
    end
    myrand2=rand;
    myrand3=rand;
    if myrand2<PAR
       if myrand3<=0.5
          for i=1:subsetsize
             new1(i)=new1(i)-1;
             if (new1(i)==0)
                new1(i)=new1(i)+1;
             end
             if abs(new1(i))>26 || new1(i)<0
                new1(i)=round(mod(new1(i),26));
             end
          end
       else
          for i=1:subsetsize
             new1(i)=new1(i)+1;
             if (new1(i)==0)
                new1(i)=new1(i)+1;
             end
             if abs(new1(i))>26 || new1(i)<0
                new1(i)=round(mod(new1(i),26));
             end
          end
       end
    else
    end
    new1(subsetsize+1)=27;



    if (length( unique(new1(1,1:subsetsize)))==subsetsize)


    a='Mt';
    b=num2str(i);
    ankara=strcat(a,b);
    for m=1:(subsetsize+1)
       mynames{m}=featureNames{populasyon(i,m)};
    end

%   POPtrial(i,1:6);
    opop=mdata(:,new1(1,1:(subsetsize+1)));
    yy=matlab2weka(ankara,mynames,opop);
    wekaClassifier = trainWekaClassifier(yy,'trees.J48',{'-D'});
    % wekaClassifier = trainWekaClassifier(yy,'bayes.NaiveBayes',{'-D'});
    predicted=wekaClassify(yy,wekaClassifier);
    actual = yy.attributeToDoubleArray(subsetsize); %java indexes from 0
    jjj=[predicted actual abs(predicted-actual)];
    h=length(find(jjj(:,3)==1));
```

```
    new1(1,subsetsize+2)=100*(h/178);
    [worst,worstindex]=max(populasyon(:,(subsetsize+2)));


          if populasyon(worstindex,(subsetsize+1))>new1(subsetsize+2)
             populasyon(worstindex,:)=new1(:);
          end
       [bestingen(iteration),indeksim]=min(populasyon(:,(subsetsize+2)));
    else
       hh=hh+1;
       deneme=randperm(26);
       new1(1,1:subsetsize)=deneme(1:subsetsize);
       new1(subsetsize+1)=27;
       a='Mt';
    b=num2str(i);
    ankara=strcat(a,b);
    for m=1:(subsetsize+1)
       mynames{m}=featureNames{populasyon(i,m)};
    end

%    POPtrial(i,1:6);
    opop=mdata(:,new1(1,1:(subsetsize+1)));
    yy=matlab2weka(ankara,mynames,opop);
    wekaClassifier = trainWekaClassifier(yy,'trees.J48',{'-D'});
    % wekaClassifier = trainWekaClassifier(yy,'bayes.NaiveBayes',{'-D'});
    predicted=wekaClassify(yy,wekaClassifier);
    actual = yy.attributeToDoubleArray(subsetsize); %java indexes from 0
    jjj=[predicted actual abs(predicted-actual)];
    h=length(find(jjj(:,3)==1));
    new1(1,(subsetsize+2))=100*(h/178);
    [worst,worstindex]=max(populasyon(:,(subsetsize+2)));


          if populasyon(worstindex,(subsetsize+2))>new1((subsetsize+2))
             populasyon(worstindex,:)=new1(:);
          end
             [bestingen(iteration),indeksim]=min(populasyon(:,(subsetsize+2)));
    end
end




 %          if bestingen(iteration<0.05)
%          %           end
```

7. UYGULAMA VE SONUÇLAR

7.1 Bulguların Değerlendirilmesi

Veri madenciliği uygulaması gerçekleştirileceği zaman, öncelikle eldeki veri ve çözülmesi gereken problem çok iyi bir şekilde analiz edilmeli ve anlaşılmalıdır. Bu iki durum, veri madenciliği uygulamasının başarı oranını etkileyecek en temel unsurları oluşturmaktadır. Eldeki veri ve problem anlaşıldıktan sonra, diğer bir önemli nokta ise uygulanacak veri madenciliği tekniğinin amaca uygun bir şekilde doğru olarak seçilmesidir.

Elde ettiğimiz veride eksiklikler, hatalar ve uyumsuzluklar olabilir. Bu yüzden uygulamaya başlarken ilk olarak eldeki veri ön işlemlerden geçirilmeli bu sayede veri seti hatalardan ve sorunlardan ayıklanmalı ve seçilmiş olan tekniğin gereksinimlerine uygun hale getirilmelidir.

Veri madenciliği uygulamalarının diğer bir can alıcı noktası ise, ön işlemlerden geçirilmiş olan veriye, amaca uygun olarak seçilmiş tekniğin uygulanmasıyla elde

edilen örüntülerin değerlendirilmesi ve yorumlanmasıdır. Elde edilen örüntünün, konunun uzmanı tarafından yorumlanması ile anlamlı bilgi elde edilmiş olur.

Veri madenciliği uygulamasının son aşaması ise elde edilen bilginin kullanıcıların anlayabileceği şekilde görselleştirilmesidir.

Tıp dünyasında veri madenciliği yaklaşımlarıyla hastalara hizmet sunulması gün geçtikçe artan konulardan olmaya başlamıştır. Bu amaca yönelik veri madenciliğinde yaygın olarak kullanılan öznitelik seçme algoritmarının yetersiz kaldığı durumlarda veya seçilen özniteliklerin uygunluğunun test edilmesi gerekn durumlarda kullanılması için tıbbi veri üzerinde daha önce denenmemiş bu yeni yaklaşımla en iyileme yöntem tabanlı yeni bir öznitelik seçme algoritması geliştirerek hekimlere zaman ve kaynak tasarrufu sağlayacak karar destek mekanizması oluşturulmuştur.

Korelasyon tabanlı öznitelik seçme algoritması, bilgi kazanç ve kazanç oranı algoritmaları kullanılarak elde edilen alt kümelerle harmoni arama algoritması kullanılarak elde edilen öznitelikler hemen hemen birbirleriyle benzeşmektedir.

WEKA kullanılarak elde edilen 5'li, 6'lı alt kümelerle; Matlab de Harmoni arama algoritması çalıştırılarak elde edilen sonuçlar aşağıdaki tabloda gösterilmektedir.

Çizelge 7.1: Algoritma Karşılaştırılması

	Bilgi Kazanç	**Kazanç Oranı**	**CFS**	**HS**
KAK	Age	Age	Age	Age
	ToT	Diabetus Melitus	ToT	Sex
	Diabetus Melitus	ToT	Diabetus Melitus	ToT
	Rose Anjina	P	Rose Anjina	Ca
	P	Rose Anjina	Donor_Type	P
	Donor_Type	Past Cardiac Disease	P	Ca_P_Product

Bu çalışmada Bilgi Kazanç ve Kazanç Oranı algoritmalarıyla Korellasyon Tabanlı Öznitelik seçme algoritmalarının hemen hemen birbirleriyle benzer sonuçlar verdikleri gözlemlenmiştir

Bu çalışmada algoritmanın başarım oranını test etmek için UCI'dan almış olduğumuz hepatit veri setiyle hazırlanan algoritmanın sonuçlarını MATLAB'de kontrol edilmiştir. Bu sonuçlar aşağıdaki tabloda gösterildiği gibidir

Çizelge 8.1: Algoritma Sonuçları

	Bilgi Kazanç	**Kazanç Oranı**	**CFS**	**HS**
KAK	Age	Age	Age	Age
	ToT	Diabetus Melitus	ToT	Sex
	Diabetus Melitus	ToT	Diabetus Melitus	ToT
	Rose Anjina	P	Rose Anjina	Ca
	P	Rose Anjina	Donor_Type	P
	Donor_Type	Past Cardiac Disease	P	Ca_P_Product
Hepatit	Albumin	Ascites	Ascites	Sex
	Bilirubin	Bilirubin	Albimun	Fatigue
	Ascites	Albimun	Bilirubin	Liverbig
	Spiders	Varices	Spiders	Ascites
	Fatigue	Spiders	Protime	Varices
	Histology	Fatigue	Varices	Histology
	Malaise	Histology	Histology	

Bu çalışmada, sadece öznitelik seçme algoritmaları ve en iyileme yöntem tabanlı öznitelik seçme algoritmalarına odaklanılmıştır. Çalışmanın birsonraki aşamasında açık kaynak kodlu veri madenciliği programı olan WEKA'nın sonuçlarına da bağlı kalmayıp Rapid Miner, SPSS Clementine gibi diğer veri madenciliği programlarıyla aynı veri setleriyle testler gerçekleştirilip paralel sonuçların birbirlerini tutup tutmadağı kontrol analizlerini de içerecek şekilde geliştirilmesi önerilir.

KAYNAKLAR

1. Kaya H., "Veri Madenciliği Kavramı Ve Uygulama Alanları", 2008.

2. Seifert J. W., Data Mining: An overview, CRS Report for congress, The Library of Congress,2004.

3. Tiryakiler I., Seyahi N., Albayrak S., Sayın K.E., Ergin E., Dağ H., "Veri Madenciliği Teknikleri ile Böbrek Nakli Geçirmiş Hastalarda koroner Arter Kalsifikasyonunun İncelenmesi", Inista 2011.

4. Kaya, E., Bulun, M., Arslan, A., "Tıpta Veri Ambarları Oluşturma ve Veri Madenciliği Uygulamaları" ,***AKADEMİK BİLİŞİM 2003, Çukurova Üniversitesi,*** Adana,(2003)

5. Prof. Dr. Akpınar, H., "Veri Tabanlarında Bilgi Keşfi ve Veri Madenciliği ", ***İ.ÜİşletmeFakültesi Dergisi,*** 29(1):1-22(2000)

6. Chakrabarti, S., Witten, I. et al., " Data Mining Know It All ", ***Morgan Kaufmann***, 2,14,39,40,41,108 (2008)

7. Han, J., Kamber, M., " Data Mining Concepts and Techniques 2nd Ed.", Editor : Jim Grey, The Morgan Kaufmann Series in Data Management Systems, Morgan Kaufmann 2,8,12,14,15,29,30,398-403, (2006)

8. Baykasoğlu, A., "Veri Madenciliği ve Çimento Sektöründe Bir Uygulama", Akademik Bilişim 2005, Gaziantep, 171:14, (2005)

9. Nong, Ye, " Part II Management of Data Mining Chapter 1 Data Collection, Preparation, Quality and Visualization ", The Handbook of Data Mining, ***Lawrence Erlbaum Associates***, 368,391,395,396,406 (2003)

10. Elder, John F., Abbott, D. W., " A Comparison of Leading Data Mining Tools", ***4th Annual Conference on Knowledge Discovery & Data Mining***, New York, USA, 234 (1998)

11. *Weka: Data* Mining Software in Java, http://www.cs.waikato.ac.nz/ml/weka/.

12. Cios, K.J., Moore G.W., "Uniqueness of Medical Data Mining", ***Artificial Intelligence in Medicine journal***, 26(1-2):1-24 (2002)

13. Karaolis M. A., Moutiris J. A., Hadjipanayi D. ve Pattichis C. S., Assessment of the Risk Factors of Coronary Heart Events Based on Data Mining With Decision Trees, IEEE Transactions On Information Technology In Biomedicine, Vol. 14, No. 3, May 2010.

14. Cheng T.H., Wei C.P., Tseng V.S., *Feature Selection for Medical Data Mining: Comparison of Expert Judgement and Automatic Approaches*, 19th IEEE Symposium on Computer-Based Medical Systems 2006.

15. Kirkpatrick S., Gelatt C. D., Jr.,Vecchi M. P., Optimization by Simulated Annealing, 13 May 1983, Volume 220, Number 4598

16. Goldberg D. E., Genetic algorithms in search optimization and machine learning, 1989, Addison Wesley.

17. Kennedy J., Eberhart R., Particle Swarm Optimization, Proceedings of IEEE International Conference on Neural Networks IV pp. 1942-1948. doi:10.1109/ICNN.1995.488968.

18. Geem Z. W., Kim J. H. and Loganathan G.V., A New Heuristic Optimization Algorithm: Harmony Search, Simulation, February 2001: 76: 60-68.

19. Srinivas K., Rao G.R., Govardhan A.,*Analysis of Coronary Heart Disease and Prediction of Heart Attack in Coal Mining Regions Using Data Mining Techniques*, The 5th International Conference on Computer Science & Education Hefei, China. August 24–27, 2010, 1344 .

20. Kantardzic, M., "Chapter 9: Artificial Neural Networks Chapter 1-1.4", Data Mining Concepts, Models, Methods and Algorithms, ***John Wiley & Sons***, (2003)

21. Dr. Özkan, Y., "Veri Madenciliği Yöntemleri",2008

22. Yıldırım, S., " Tümevarım Öğrenme Tekniklerinden C4.5'in İncelenmesi", Yüksek Lisans Tezi, ***İstanbul Teknik Üniversitesi Fen Bilimleri Enstitüsü*** , İstanbul, (2003)

23. J. H. Gennari, P. Langley ve D. Fisher. Models of incremental concept. *Artificial Intelligence*. 1989, 40

24. Kennedy, J. and Eberhart, R. C. "Particle swarm optimization" Proc. IEEE int'l conf. on neural networks Vol. IV, pp. 1942-1948. IEEE service center, Piscataway, NJ, 1995.

25. T. Seçkin, K. Cihan 'Parçacık Sürüsü Optimizasyon Algoritması ve Benzetim Örnekleri'

26. Geem ZW, Kim JH, Loganathan GV., "A new heuristic optimization algorithm: harmony search" Simulation, Vol.76, No.2, s.60-68, 2001.

27. Ceylan O. ," Akıllı Yöntem Tabanlı Tekli ve İkili Kısıtlılık Analizi" ,2012

28. Seyahi, N., Kahveci, A., Cebi, D. Altiparmak, M. R., Akman, C., Uslu, I., Ataman, R., Tasci, H., ve Serdengecti, K., "Coronary artery calcification and coronary ischaemia in renal transplant recipients", *Nephrol Dial Transplant*, 26 (2):720-6, 2011.

ÖZGEÇMİŞ

İsim Soyisim: Çağıl Acar Şaylan

Doğum Yeri ve Tarihi: İzmir, 05 Şubat 1987

Eğitim: Kadir Has Üniversitesi, İstatistik ve Bilgisayar Bilimleri ve Bilgisayar Mühendisliği Bölümü, 2009

Kadir Has Üniversitesi Yönetim Bilişim Sistemleri Yüksek Lisans Programı 2013

Yayınlar:

Bildiriler:

- Oğuzhan Ceylan, **Çağıl Acar Şaylan**, Işıl Yenidoğan, Hasan Dağ, Intelligent Method on Feature Selection Algorithms for Medical Data Sets, IX. Tıp Bilişimi Kongresi, Antalya, Türkiye, 15-17 Kasım 2012.
- Hasan Dağ, Kamran Emre Sayın, Işıl Yenidoğan, Songul Albayrak, **Çağıl Acar**, Comparison of Feature Selection Algorithms for Medical Data, International Symposium on Innovations in Intelligent Systems and Application INISTA 2012, Trabzon, Türkiye, 2-4 Temmuz 2012.