GENETİK VE GENETİK OLMAYAN FAKTÖRLERE BAĞLI OLARAK TÜRK HASTALARDA VARFARİN DOZAJINI TAHMİN EDEN BİR UZMAN SİSTEM GELİŞTİRİLMESİ

Osman ALTAY

Yüksek Lisans Tezi
Yazılım Mühendisliği Anabilim Dalı
Danışman: Yrd. Doç. Dr. Mustafa ULAŞ
ŞUBAT-2016

T.C.
FIRAT ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

GENETİK VE GENETİK OLMAYAN FAKTÖRLERE BAĞLI OLARAK TÜRK HASTALARDA VARFARİN DOZAJINI TAHMİN EDEN BİR UZMAN SİSTEM GELİŞTİRİLMESİ

YÜKSEK LİSANS TEZİ

Osman ALTAY

(141137113)

Tezin Enstitüye Verildiği Tarih : 20 Ocak 2016
Tezin Savunulduğu Tarih : 05 Şubat 2016

Tez Danışmanı : Yrd. Doç. Dr. Mustafa ULAŞ (F.Ü.)
Diğer Jüri Üyeleri : Yrd. Doç. Dr. Oğuz ATA (İ.A.Ü.)
Yrd. Doç. Dr. Murat KARABATAK (F.Ü.)

ŞUBAT-2016

ÖNSÖZ

Bu Yüksek Lisans Tezi çalışmasında, genetik ve genetik olmayan faktörlere bağlı olarak Türk hastalarda varfarin dozajını tahmin eden bir uzman sistem hazırlanmıştır. Biyoenformatik alanı ve veri madenciliği algoritmalarından Bayesyen ve K-en yakın komşu algoritmaları incelenmiştir.

Tez çalışması sürecinde yardımlarını, desteklerini ve fikirlerini hiçbir zaman esirgemeyen danışman hocam Sayın Yrd. Doç. Dr. Mustafa ULAŞ'a en içten teşekkürlerimi sunuyorum.

Mahmut ÖZER'e "*The Effect Of Polymorphisms In Cytochrome P450 2C9, Cytochrome P450 4F2, Epoxide Hydrolase 1 And Vitamin K Epoxide Reductase 1 On Warfarin Dose In Turkish Patients*" isimli Yeditepe Üniversitesi'nde 2011 yılında sunduğu yüksek lisans tezi sonucunda elde ettiği verileri kullanmama izin verdiği için en içten teşekkürlerimi sunuyorum.

Öğrenim hayatımda olduğu gibi, tez yazım sürecinde hiçbir fedakârlıktan kaçınmayan aileme ve dostlarıma candan teşekkürlerimi sunuyorum.

Osman ALTAY

ELAZIĞ-2016

İÇİNDEKİLER

ÖZET

Genetik mühendisliğinin önemi son yıllarda hızla artmaktadır. DNA hakkındaki verilerin büyük çoğunluğu 90'lı yıllardan sonra ortaya çıkarılmıştır. Gen alanındaki çalışmalar gün geçtikçe artmakta ve bu çalışmalardan büyük miktarda analiz gerektiren veriler elde edilmektedir. Gen araştırmalarından elde edilen ve hızla artan bu veri yığınlarının anlamlandırılması ile birlikte insanoğlunun yaşam kalitesini arttırmaya yönelik birçok çalışma yapılmaktadır. Bu bağlamda, ilaç kullanımını en iyi seviyede tutarak insan ömrünü uzatmak önem verilen çalışmalar arasında yer almaktadır.

Bu tez çalışmasında biyoenformatik, genom kavramları ve uzman sistemler incelenmiştir. Türk hastalar üzerinden elde edilen veriler üzerinde, veri madenciliği algoritmalarından K-en yakın komşu ve Bayesyen algoritmaları uygulanarak varfarin kullanımı en iyi seviyede tutulmaya çalışılmıştır.

Anahtar Kelimeler: Genetik Mühendisliği, Biyoenformatik, Veri Madenciliği, K-En Yakın Komşu, Bayesyen Algoritması

SUMMARY

Improvement of an Expert System That Predict Warfarin Dosage in Turkish Patients Depending on Genetic and Non-Genetic Factors

The importance of genetic engineering has been increasing rapidly in recent years. The vast majority of data about DNA was discovered after 90's. The studies on field gene is increasing day by day and large amounts of data are obtained from these studies that requires analysis. Many studies are done to improve the life quality of mankind by signification of these rapidly increasing stack of data obtained from gene investigations. In this context, extending human life by keeping the drug utilization at optimum level is involved in regarded studies.

In this thesis, the concepts of bioinformatics and genom and expert systems are examined. The usage of warfarin is tried to be kept at optimum level by applying data mining algorithms K-nearest neighbor and Bayesian algorithms on the data obtained from Turkish patients.

Key Words: Genetic Engineering, Bioinformatics, Bioinformatics Databases, Data Mining, K-Nearest Neighbor, Bayessian Algorithm

KISALTMALAR

DNA	: Deoksiribonükleik Asit
A	: Adenin
G	: Guanin
S	: Sitozin
T	: Timin
KNN	: K-Nearest Neighbor (K-En Yakın Komşu Algoritması)
VKOR	: Vitamin K Epoxide Reductase

ŞEKİLLER LİSTESİ

Sayfa No

TABLOLAR LİSTESİ

Sayfa No

SEMBOLLER LİSTESİ

$\boldsymbol{\sigma_x}$	: Standart sapma
$\boldsymbol{\overline{x}}$	: *x* değerlerinin aritmetik ortalama
$\boldsymbol{e}$	: Üstel fonksiyon
$\boldsymbol{\pi}$	: Pi sayısı
$\boldsymbol{\mu_{c_i}}$	: Belirli bir sınıftaki değerlerin aritmetik ortalaması

1. GİRİŞ

Son 60 yıl içerisinde enformatik ve biyolojik alanlarındaki hızlı ilerlemenin ışığında, yeni bir bilim dalı olan biyoenformatik ortaya çıkmıştır. Keşfedilmemiş birçok araştırma alanı olan ve ilerlemeye açık olan disiplinler arası bir bilim dalı olan biyoenformatiğin önümüzdeki yıllara damgasını vurması beklenmektedir.

İlerleyen bilgisayar teknolojisi ile beraber, genlerin daha iyi araştırılması sağlanmış ve gen teknolojisinde hızlı gelişmeler yaşanmıştır. Bilgisayar ortamında toplanan verilerin daha hızlı analiz edilmesi ve kıyaslanması, tıp, tarım, hayvancılık ve çevre gibi birçok konuda özgün ve etkili yaklaşımlar getirerek insanoğlunun hayatının daha konforlu ve sağlıklı olmasını sağlamıştır.

İnsanoğlunun hayat standartlarını daha ileriye taşıyacak biyoenformatik bilim dalında da tıpkı diğer alanlarda olduğu gibi bilişim alanındaki daha hızlı ve etkili sonuçlar sunan teknoloji ve metodolojilere gerek duyulmaktadır. Biyoenformatik alanında geliştirilen dizi kıyaslama algoritmaları ve uzman sistemlerde kullanılan algoritmalar bu konuda büyük bir öneme sahiptir. Ayrıca büyük veri çalışma alanı ile biyoinformatik veri yapısı itibariyle ilişki barındırmaktadır. İncelenecek olan verilerin boyutu ciddi anlamda sorun teşkil etmektedir. Bu bakımdan veri işleme tekniklerinin büyük veriler üzerinde uygulanmasını sağlayacak olan büyük veri işleme yöntemleri de biyoinformatik alanı için önemli çalışma alanlarındandır.

Gerçekleştirilen bu veri işleme yöntemleri ile biyolojik yapının davranış tarzının anlaşılması amaçlanmaktadır. İşlenen veriden oluşturulan bilgi birikimi ile bir insanın ilaç tepkisi veya bir tohumdan oluşacak olan bitkinin davranışı tahmin edilebilmektedir. Gereğinden fazla ya da eksik ilaç kullanımı insan sağlığını ölümle sonuçlanabilecek derecede kötü etkileyebilmektedir. Gelişen gen teknolojisi ile birçok genin insanın vücudunda hangi metabolik olayları etkilediği bilinmektedir. Varfarin kanın pıhtılaşmasını önleyici bir ilaç olmasına rağmen dozaj ayarlaması güç olan bir ilaçtır. Kanın pıhtılaşmasına etkisi olan bazı genlerin öğrenilmesi ve genetik olmayan faktörlerin de kullanılması sonucunda bu ilacın dozaj miktarının belirlenebileceği öngörülmüştür.

2. BİYOENFORMATİK VE GENETİK MÜHENDİSLİĞİ

Son yıllarda hızlı gelişen bilim dalları arasında genetik ve bilişim yerini almıştır. Özellikle genetik alanında yapılan araştırmalar ile tıp, tarım, hayvancılık ve çevre konularında önemli adımlar atılmaktadır. Genetik alanında yapılan birçok araştırma direkt olarak bilişim alanı ile ilişkilendirilebilmektedir. Her alanda olduğu gibi genetik bilimi de bilişim alanındaki metot ve yöntemleri kullanmaktadır. Bu sayede gerekli olan verilere daha hızlı ve net bir şekilde ulaşılırken, bu verilerin işlenmesi ve sonuçların elde edilmesi kolaylaşmaktadır. Bu iki bilimin bu kadar yakınlaşması sonucu biyoenformatik kavramı ortaya çıkmıştır [1, 2].

2.1. Enformatik

Farklı yöntemler vasıtasıyla çeşitli amaçlar için toplanan sayısal verilerin; depolanması, sınıflandırılması, erişilebilmesi, dağıtılması veya işlenmesi ile uğraşan uygulamalı ve kurumsal bilim dalına bilgi bilimi denilmektedir [3, 4]. Enformatik kelimesi bizi doğrudan bilişime götürmektedir. Bilişim ise insanoğlunun hayatındaki her alanda kullandığı bilginin, elektronik araçlar vasıtasıyla mantıklı ve düzenli bir şekilde işlenmesidir. Enformatik bilimi ile neredeyse bütün araştırma alanlarındaki bilgiye daha hızlı erişmek ve bilgiyi doğru bir şekilde yorumlamak mümkündür. Toplanan verilerin;

- Daha hızlı işlenmesi,
- Birbiri ile ilişkilendirilmesinin daha hızlı olması,
- Önceden belirlenmiş kurallara göre anlamlandırılması,
- Sonucunda daha hızlı ve doğru sonuç elde edilmesi,
- Sayısal olarak saklanması gibi avantajlardan dolayı enformatik kavramının bilimin ilerlemesine sağladığı katkı azımsanamayacak kadar çoktur.

Enformatik alanı özellikle tıbbi enformatik, işletme enformatiği, endüstriyel enformatik, kimyasal enformatik, sağlık enformatiği ve biyoenformatik gibi alanlarla yakın ilişki içindedir. Bu alanlara bakıldığında bilgiyi işleme yanında bilgilerin derlenmesi, analizi ve yorumlarını da kapsadığı görülmektedir [5].

2.2. Biyoenformatik

Son yıllarda geliştirilen yeni yöntem ve teknikler; genetik biliminin ilerleyişini çok önemli ölçüde arttırmıştır. Özellikle DNA dizi analizi yöntemlerinin hızlanması ve DNA hakkında daha çok bilgi elde edilmesi ile birlikte çeşitli türlerin genomlarının DNA dizilimi elde edilmiştir. Genomik bilginin bu kadar hızlı büyümesi, bilginin depolanması, ilişkilendirilip düzenlenmesi, yardımcı programlar vasıtası ile analiz edilebilmesi için bilişim alanına olan muhtaçlığını arttırmaktadır [6, 7].

Biyoenformatik biliminin etkili bir şekilde kullanılması ile bilgisayarla standart düzeyde ilişki içinde olan biyolog ve genetik mühendislerinin daha hızlı, rahat ve optimum performans ile analiz edilmiş verilerden faydalanmaları sağlanabilmektedir. Çok büyük boyutlarda genetik bilgi içeren verilerin analiz edilmesi, genetik mühendislerine çok büyük zaman kaybı yaşatmaktadır. Büyük boyutlarda olan bu genetik verilerin analizinde kullanılacak olan yöntem ve metotlar verimliliği arttırmada önemli yere sahiptir. Biyoenformatik alanının temel konusu ise metot ve yöntemlerin geliştirilmesi, verimli kullanılması ve uygulamaların yapılmasını kapsamaktadır.

Biyoenformatik alanının kullanılması ile birlikte verilerin depolanması ve yeni verilerin eklenmesi için etkin veri tabanları oluşturulabilmekte, bu verilerin analizinde kullanılan yöntem ve metotlar geliştirilmekte ve elde edilen verilerin biyolojik açıdan anlamlandırılması ve yorumlanması sağlanmaktadır [8].

Biyoenformatik alanının hedef aldığı çalışma alanları:

- Biyolojik enformasyonun paylaşımının kolaylaştırılması,
- Bilgisayar ile optimum seviye ulaştırılmış veri analizi ve iletimi,
- DNA sıra dizilim çalışmaları,
- Protein sıra dizilim çalışmaları ve fonksiyonlarının belirlenmesi,
- DNA, RNA ve proteinin üç boyutlu araştırılması,
- İnsan, hayvan ve tek hücreli canlıların genom projelerinden elde edilen verilerin depolanması, erişilmesi ve analizi,
- Biyolojik gelişimlerin simülasyonları,
- Biyolojik açıdan etkin moleküllerin araştırılması,
- Veri madenciliği ve metin madenciliği araştırmaları,

- Herhangi bir biyolojik fonksiyona tepki veren küçük moleküllerin tasarlanması,
- Karmaşık genetik fonksiyon ya da regülasyonların araştırılması,
- Tıbbi ya da endüstriyel amaçlı yeni makro ya da mikro moleküllerin araştırılması,
- Genetik faktörlerin herhangi bir hastalık üzerindeki etkilerinin araştırılması,

Yukarıdaki gibi sıralanabilecek problemlerin verilerinin toplanması, sayısal bir şekilde depolanması, yönetimi ve bu verilerin analizi için matematik ve bilgisayar bilimlerinde faydalandığı çözüm metotları aşağıda sıralanmaktadır [9].

- Arama ve desen tanıma yöntemleri,
- Yapay sinir ağları,
- Yapay zekâ ve uzman sistemler,
- Veri tabanı yönetimi,
- Genetik algoritmalar,
- Evrimsel görüntüler,
- Kümeleme algoritmaları.

2.3. Biyoenformatiğin Tarihsel Gelişimi

Biyoenformatiğin başlangıcı olarak 1950 ve 1960'lı yıllar kabul edilmektedir. Tam olarak başlangıç tarihi belirlemek zor olsa da, Pualing ve Corey'in 1951 yılında proteinlerin ikincil yapılarının tahmini için yaptıkları uygulama biyoenformatiğin tarihsel başlangıcı olarak kabul edilebilir [10]. Ancak bu alanda bilgisayar destekli ilk uygulamanın 1966 yılında Scientific American dergisinde yayınlanması vesilesi ile biyoenformatiğin gerçek başlangıcının 1966 yılında olduğunu varsaymak daha gerçekçi olacaktır. Bu yıllarda bilgisayar uygulamalarının biyolojide kullanılmaya başlanması ve her iki alanın da gelişime açık olması sebebiyle, iki alandaki gelişimlerle beraber biyoenformatik dalı günümüzde endüstriyel ve akademik alanda en popüler bilim dalları arasında yerini almıştır. Biyoenformatik terimi 1980'li yılların başlangıcında ortaya atılmış ve o yıllardan sonra kullanılmaya başlanmıştır. Biyoenformatik terimi ile beraber aynı anlamda olan moleküler biyoenformatik, computational biology, biocomputing terimleri de kullanılmaya başlanmıştır [11]. 1988 yılında National Center for Biotechnology Information (NCBI) adında bir kuruluş kurulmuştur. Bu kurum temel moleküler biyoloji ve genetik biliminin

anlaşılması, analizi ve analizlerin yorumlanmasında en etkili kurumdur. Biyoenformatik alanın gelişiminde en büyük rolü insan genom projesi oynamıştır. İnsan genom projesi 13 yıllık uluslararası bir çalışma sonucunda 3035 insan geninin belirlenmesi ve bu belirlenen genlerin biyolojik çalışmalarda kullanılacak düzeyde olmasını hedef almıştır [12].

3. GENOM KAVRAMI

Bütün organizmaların ve bazı virüs çeşitlerinin canlılık işlevlerini ve biyolojik gelişimlerini sürdürebilmek için gerekli olan talimatları taşıyan nükleik asite DNA (Deoksiribo Nükleik Asit) denir [13]. Bir insan hücresinde 46 kromozomun içinde 3 milyar baz çift içeren yaklaşık olarak iki metre uzunluğunda DNA bulunmaktadır. Bütün genetik bilgiyi içeren kromozom setinin tamamına genom denir [14].

Genomdaki bu bilgiler canlıyı diğer türlerden ayıran özellikleri ve kendi türündeki canlılardan ayıran boy, kilo, göz rengi, vücut yapısı, ten rengi gibi özelliklerinin yapısının yanında hastalıklara karşı direnci, metabolizma işleyişi, ilaç etkileri ve kalıtımsal olarak yakalanabileceği hastalıkların belirlenmesinde önemli rol oynamaktadır.

3.1. Genomik

Farklı türlere ait genom yapılarını inceleyen bu genom yapılarındaki kromozomlara dizilenme teknikleri uygulayan, genomların tüm yapısal ve işlevsel yönlerini inceleyen bilim dalıdır. Genomik dalının amaçları arasında bir canlı türünün bütün DNA yapısının belirlenebilmesi de vardır. Genomik bilimi bu amaçla insan genomunun yapısını, bileşimini ve geçirdiği evrimleri inceleyerek, biyolojik bir anlamı olabilecek DNA'yı tanımlamaya çalışmaktadır.

3.2. Gen

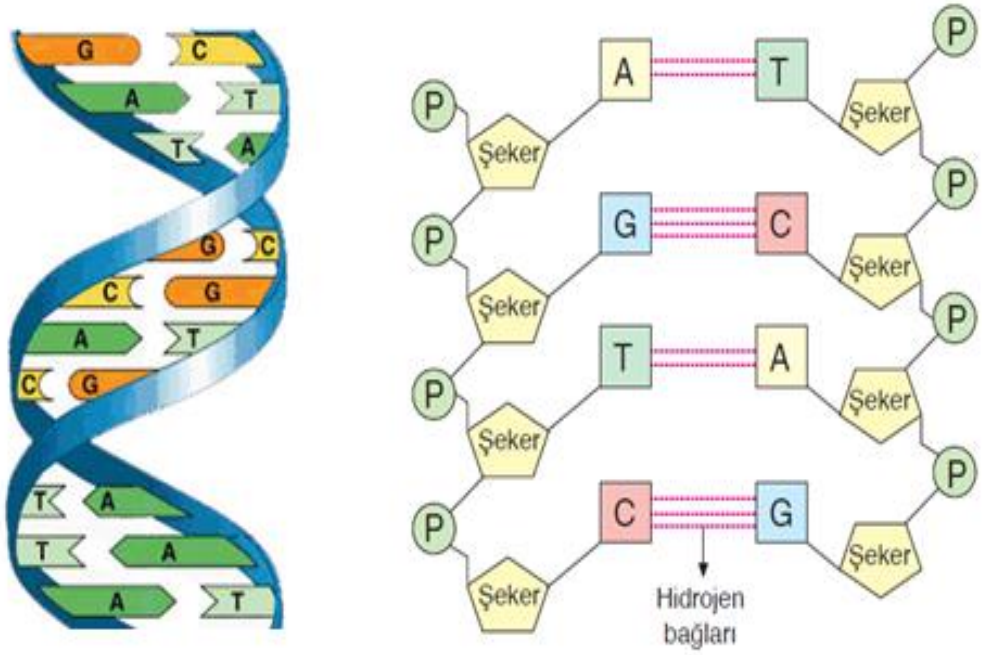
DNA'nın bir parçası olan genler organizmaların özelliklerini belirlemektedir. Organizmalar içerisinden insanı ele alacak olursak genler, insanları ayıran fiziksel özellikler, insanların vücudunda hangi olayların gerçekleştiğini ve hangi hastalıkları geçirmeye eğilimli olduklarını belirlemektedir. Günümüzde insan gen araştırmaları büyük önem kazanmış ve bu alanda birçok çalışma yapılmaktadır.

Organizmalar kendisini belirleyen tüm özelliklerini yani genom yapılarını ailelerinden alırlar. "Kalıtım", anne ve babadan olan özelliklerin sonraki kuşağa geçmesi olarak tanımlanmaktadır. Bu aktarım genler aracılığı ile gerçekleşmektedir. Her genin kodladığı

bir protein bulunmaktadır. Proteinler insan vücudundaki işlevleri yerine getiren moleküllerdir. Haberleri taşırlar, reaksiyonları katalizler, hücrelerimizin içinde yer almaktadırlar [15] .

Proteinler aminoasit adı verilen yapı taşlarından oluşmuşlardır ve bu aminoasitlerin ne şekilde dizilerek protein oluşturacaklarının bilgisi DNA'da bulunmaktadır. DNA'da da bu bilgi genler aracılığıyla bulunmaktadır.

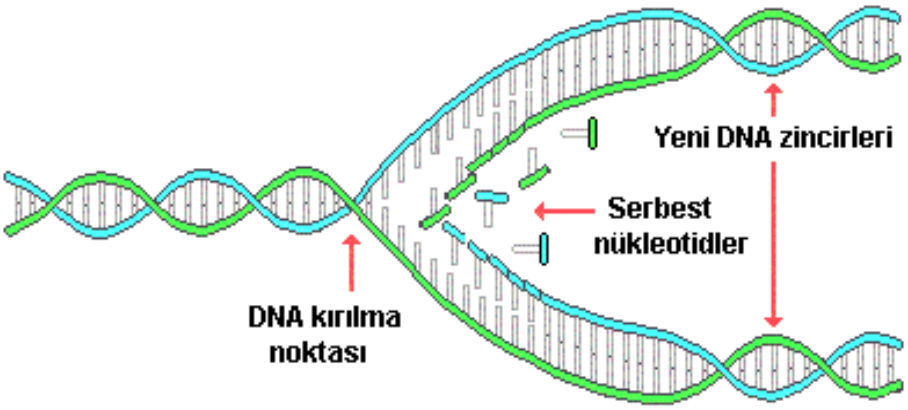
Genin yapısı kasa şifresine benzemektedir yani yan yana gelen harflerden oluşmaktadır. Bu harfler dört adet bazı tanılamaktadır. Bu bazlar Adenin (A), Guanin (G), Sitozin (C) ve Timin (T) olarak adlandırılmaktadır. Şekil 3.1'de DNA'nın genel yapısı gösterilmiştir.



Şekil 3.1 DNA'nın genel yapısı [16]

Ökaryot canlılarda DNA sentezi kendisini mayoz ya da mitoz bölünmeye hazırlayan hücrede, hücre siklusunun sentez fazı sırasında gerçekleşmektedir. DNA sentezi yarı saklı bir model ile açıklanır. DNA sarmalındaki iki sarmalın her bir ipliğinin kalıp görevi görerek kendine eş yeni bir DNA oluşturmasıdır. Yeni oluşan molekül kendisini oluşturan molekülün bir sarmalını taşıyacaktır. 1957 yılında ökaryotlarda bu replikasyon olayının yarı saklı olduğunu J.H. Taylor, P. Woods ve W.Hughes kanıtlamıştır [17]. Replikasyon biyomolekülün kendisine benzer yeni bir biyomolekül oluşturmasıdır. DNA replikasyon yapabilen tek biyomoleküldür. DNA bu özelliği ile replikasyon yaparak kendisinde taşınan

genetik bilgileri nesilden nesile aktarabilmektedir. DNA replikasyon işlemi Şekil 3.2'de gösterilmiştir.



Şekil 3.2. DNA sentezi [18]

3.3. Genetik Mühendisliğinin Çalışma Alanları

Genetik mühendisliğindeki çalışmalar sağlık başta olmak üzere birçok alanda karşımıza çıkmaktadır. Genetik mühendisliğinin çalışma alanlarından bazıları aşağıda alt başlıklar halinde verilmiştir.

3.3.1. Kişiye Özel Hekimlik

İnsanların, uygulanan ilaç tedavilerine verdiği reaksiyonlar kişiden kişiye değişmektedir. Bütün tedavilerde olmasa da, tedavilerin bazılarının kişiye yönelik olması önemli rol oynamaktadır [19].

3.3.2. İlaç Üretimi ve Geliştirilmesi

Hastalıkların gelişme süreçlerinin biyoenformatik yöntemler kullanılarak belirlenmesinin ışığında yeni etken proteinler tespit edilebilecektir. Böylece var olan hastalıklara yönelik kullanılan ilaçlar geliştirilebilecek ya da daha etkili yeni ilaçlar üretilebilecektir [20].

3.3.3. Biyoteknoloji

Biyoenformatik alanının en geniş kapsamlı başlıkları arasında yer almaktadır. Biyoteknoloji genel olarak insanların yaşamını kolaylaştırmayı, üretim gibi yeteneklerinin gen bazında araştırarak diğer canlı organizmalar üzerinde uygulanmasını amaçlamaktadır [21].

3.3.4. Antibiyotiklere Karşı Hastalıkların Direnci

Hastalıklara sebep olan mikropların mevcut antibiyotik direnci ile alakalı genlerin araştırılması yapılmaktadır. Bu genlerin tespiti sonucunda, elde edilen ilgili gen kodları antibiyotik direnci dışındaki durumlar içinde kullanılabilirliği konusunu da araştırma alanı olarak kullanmaktadır [22].

3.3.5. Mikrobiyal Gen Uygulamaları

Dünya üzerinde milyarlarca çeşit mikroorganizmalar yer almaktadır. Bu organizmalar hemen hemen her yerde yaşamlarına devam edebilmektedir. Yüksek sıcaklık ve soğuklar, radyasyona maruz kalan ortamlar, asit, tuz ve basıncın yüksek olduğu ortamlar gibi başka organizmaların yaşamlarına devam edemeyeceği alanlarda bulanabilmektedirler. Mikroorganizmaların sahip oldukları bu dayanım özellikleri direkt olarak genlerle alakalıdır. Bu mikroorganizmaların gen haritalarının incelenmesi ve ilgili genlerin tespiti ile insan hayatını oldukça kolaylaştıracak, endüstriyel uygulamalara kaynak teşkil edecek önemli bilgiler elde edilebilecektir [23].

3.3.6. Gen Terapisi

Gen terapisinin ilerleyen yıllarda en çok önem arz eden araştırmalar arasında yer alması beklenmektedir. Bu yöntemin yakın zamanda insan genleri üzerinde uygulanarak tedavi edilmesi amaçlanmaktadır. Kansere yol açan genlerin tespit edilmesi ve bu genlerin değiştirilerek kanserin tedavi edilmesi, bu alandaki en büyük başarı olacaktır [24].

3.3.7. Moleküler Hekimlik

İnsanlara ait gen haritaları üzerinde araştırmalar yaparak, bu araştırmalar sonucunda elde edilen bilgilerin moleküler bazda analiz ve karşılaştırılması yapılarak oluşturulan moleküler bilginin en iyi tedavi yönteminde kullanılması üzerinde çalışılmaktadır.

3.3.8. Adli Analizler

Adli olarak gerçekleşen olayların bazılarının gün ışığına ulaşmasında biyolojik alanda yapılan çalışmalara ihtiyaç duyulmaktadır. Bazı saldırı ve yaralamalarda saldırganın gen analizi yapılarak olaylar aydınlatılmıştır [25].

3.3.9. Biyolojik Silahlar

İnsan da dâhil olmak üzere canlılara yönelik olarak bilinen geleneksel savaş yöntemleri kullanmadan biyolojik silahlarla çok daha etkili ve büyük saldırılar gerçekleştirilmektedir. Gen haritası bilinen insanları hedef alacak özel biyolojik silah niteliğinde mikroorganizmalar geliştirilebilir [26].

3.3.10. Atıkların Zararsız Şekilde İmha Edilmesi

Kimyasal toksinler, radyasyon yayan kimyasal atıkların yok edilmesinde bazı mikroorganizmalar önemli rol oynamaktadır. Bu mikroorganizmaların incelenerek yeni türlerin geliştirilmesi amaçlanmaktadır.

3.3.11. Evrimsel Çalışmalar

Evrim Teorisini ele alarak, canlıların DNA ve gen haritalarından faydalanılarak canlılar üzerinde değişim ve etkileşimleri incelenmektedir. Evrimsel çalışmalar kullanılarak canlıların birbiri ile akrabalık ilişkileri belirlenmeye çalışılmaktadır [27, 28].

4. UZMAN SİSTEMLER

Yapay zeka alanının bir dalı olan uzman sistemler 1960'lı yılların ortasında yapay zeka topluluğu tarafından geliştirilmiştir. Uzman sistemlerin en basit açıklama şekli, uzman bir insanın yapabileceği bir işi bilgisayara aktarmaktır. Uzman sistemlerin çalışabilmesi için alanında uzman kişiler tarafından elde edilen bilgiler bilgisayar ortamında belirli kurallara göre saklanır. Bilgisayarın, çeşitli algoritmalar kullanarak bu bilgilerden çıkarım yapmasını ya da hedeflenen bir sonuca ulaşmasını sağlar. Uzman sistemler, geleneksel yöntemler ve direkt sonuca ulaşılamayacak problemleri güçlü ve esnek algoritmalar yardımı ile sonuca ulaştırır [29].

4.1. Kural Tabanlı Uzman Sistemler

Kural tabanlı uzman sistemler bilgiye ulaşmak için en temel ve basit yoldur. Kural tabanlı uzman sistemlerinin temelinde "if" yapısı yer almaktadır. Kural tabanlı uzman sistemlerde bilgi belirli kurallara göre bilgisayara verilmektedir [30]. Kural tabanlı uzman sistemler üretim planlama, eğitim, DNA histogramı, elektronik güç planlama, dolandırıcılık tespiti, sistem geliştirme, arıza teşhisi, tarım planlaması, ders sistemi, ders dağılımı ve sensor kontrolü gibi bir çok alanda kullanılmaktadır.

4.2. Bilgi Tabanlı Uzman Sistemler

Bilgi tabanlı uzman sistemler insan odaklı bir sistemdir. Yapay zeka algoritmalarını temel alarak, bilgisayar sistemine insanın bilgisini anlatmayı ve kullanmayı amaçlar [31]. Bilgi tabanlı uzman sistemlerde dört temel unsur vardır. Bunlar; bilgi bankası, çıkarım yapma motoru, bilgi mühendislik aracı ve spesifik kullanıcı ara yüzüdür [32]. Bilgi tabanlı uzman sistem, vaka doğrulama, finansal analiz, kimyasal olay yönetimi, karar destek, üretim yönetimi, tedavi planlanması, iklim tahmin ve kimyasal olay yönetimi gibi birçok alanda kullanılmaktadır.

4.3. Yapay Sinir Ağları

Yapay sinir ağları, biyolojik sinir ağları temel alınarak tasarlanmıştır. Yapay sinir ağlarında bilgi belli formüllere göre işlenmektedir. Yapay sinir ağları içerisinde bulunan yapay sinir hücreleri birbirleriyle bağlanırlar ve bu bağlantılar arasında belirli ağırlıklar ve ateşleme fonksiyonları yer almaktadır [33]. Yapay sinir ağları parametre belirleme, makine öğrenmesi, bilgi öğrenmesi, optimal güç belirlenmesi, işlem denetimi ve biyomedikal gibi birçok alanda kullanılmaktadır.

4.4. Bulanık Mantık

Bulanık mantık, bilgi kümesi sonucunda kesin ve net sonucu olmayan çıkarımlar elde ediliyorsa kullanılmaktadır. Bulanık mantık, matematiksel işlemler kullanılarak, geleneksel bilgisayar tekniklerine göre daha net sonuçlar vermese de insan tepkilerine en yakın sonuçları vermektedir [34]. Bulanık mantık, arıza tespiti, performans belirleme, bilgisayar güvenliği, kontrol sistemleri ve tıbbi teşhis gibi birçok alanda kullanılmaktadır.

5. VERİ MADENCİLİĞİ

Büyük boyutlardaki anlamsız verilerin içerisinden algoritmalar ve belirli kuralları kullanarak, önceden tahmin edilemeyen ancak kullanım olarak faydalı bilgilerin ortaya çıkarabilmesi için verileri analiz etme ve inceleme sürecine veri madenciliği denilmektedir [30].

Veri madenciliği (Data Mining), özellikle tıp, elektronik ticaret, bilim, iş ve eğitim alanlarındaki uygulamalara temel oluşturabilecek bir araştırma sahası olarak ortaya çıkmıştır.

Veri madenciliği genel olarak elde bulunan anlamsız bilgilerden, veri madenciliği yöntemleri kullanılarak anlamlı ve faydalı yeni bilgileri elde etme işlemidir. Veri madenciliği, bilgi keşfi sürecinin içerisinde yer almaktadır. Veri madenciliği işlemi yapılırken aşağıdaki adımlar uygulanmaktadır [30].

1. *Veri Temizlenmesi:* Elde edilen veri tabanı içerisindeki verilerden anlamsız olanların çıkarılması.
2. *Veri Entegrasyonu:* Farklı verilerin anlamlı bir şekilde birleştirilmesi.
3. *Veri Seçimi:* Anlamlı verilerin elde edilmesi için kullanılacak verilerin belirlenmesi.
4. *Veri Transferi:* Veri madenciliği algoritmaları uygulanacak verilerin, kullanılması tasarlanan veri madenciliği algoritması için uygun hale getirilmesi.
5. *Veri Madenciliği:* Veri madenciliği algoritmaları için uygun hale getirilmiş veri tabanına, veri madenciliği algoritmasının uygulanması.
6. *Veri Değerlendirmesi:* Kullanılan algoritma sonucunda elde edilen örüntü içerisinde farklı örüntülerin tespit edilmesi.
7. *Veri Sunumu:* Veri madenciliği algoritmaları sonucunda elde edilen yeni bilgilerin kullanıcıya sunulması.

5.1. Veri Madenciliğinin Günümüzdeki Yeri

Veri madenciliği, bilgi endüstrisi içerisinde yer alan büyük ve önemli miktardaki bilgiyi, anlamlı ve faydalı bir şekilde kullanılabilir hale getirdiği için büyük önem taşımaktadır. Kazanılan bilgi ile market analizleri, dolandırıcılık tespitleri, müşteri tutma, üretim kontrolü ve bilim keşiflerinde önemli kazançlar elde edilmiştir [30]. Bu faydaları sayesinde veri madenciliği gün geçtikçe önemini arttırmakta ve kullanım alanların çokluğu ve yaygınlığı sayesinde her geçen gün gelişmektedir.

Veri madenciliği bilgi teknolojisinin doğal gelişiminin bir sonucu olarak da görülebilir [30]. Veri tabanı ve veri tabanı oluşturma işlemlerinin gelişmesi ile birlikte verilerin depolanması, kullanılması ve işlenmesi önemli bir rol almıştır. Bütün bunların gelişimi sürecinde doğal olarak veri analizi önemli bir rol oynamakta ve ilerleyen zamanda önemini arttırarak devam etmesi beklenmektedir. Önemini arttırmasının en büyük örneği ise 1960'dan bu yana başlayan ilkel veri depolama ve işleme sistemleri günümüzde sistematik olarak gelişmiş güçlü veri işleme ve veri tabanı sistemlerine yerine bırakmıştır.

5.2. Veri Madenciliğinde Bilginin Keşfi ve Adımları

Büyük boyutlardaki veriler veri tabanlarında tutulmakta ve sorgulama işlemleri sonucunda kullanıcının önüne gelmektedir. Veri tabanlarında bulunan bu büyük boyutlardaki verilerin anlamlı hale gelmesi için veri madenciliği algoritmaları uygulanmaktadır. Bu anlamlı bilgilerin çıkarılması işlemine ise veri tabanında bilginin keşfi denmektedir. Büyük boyutlarda depolanan bu bilgilerden anlamlı bilgilerin çıkarılması işlemi yeni geliştirilen tekniklerle yapılmaktadır fakat bazen bu teknikler de yetersiz kalmaktadır. Bu yüzden veri madenciliği çalışmaları günümüzde güncelliğini korumaktadır. Prof. Dr. Usama Fayyad'a göre veri tabanında bilgi keşfi sürecinde aşağıdaki adımlar izlenmelidir [35].

1. *Veri Seçimi:* Veri tabanı içerisinde yer alan veri kümeleri içerisinden kullanılması uygun olacak verilerin seçilmesi işlemidir. Bu seçilme işlemi sonucunda elde edilen yeni verilere örneklem kümesi denmektedir.
2. *Veri Temizleme ve Ön İşleme:* Veri seçimi sonucunda elde edilen veriler içerisinden veri madenciliği algoritması için faydasız olacak verilerin seçilmesi bu

verilerin çıkarılması veya değiştirilmesi işlemidir. Bu adım sayesinde veri madenciliği teknikleri veri kümesi üzerinde daha en iyi şekilde çalışarak daha doğru sonuçlar bulabilecektir.

3. *Veri Madenciliği:* İlk iki adımın gerçekleştirilmesi sonucunda elde edilen veri kümesine veri madenciliği tekniklerinin uygulanması işlemidir. Bu veri madenciliği algoritmaları kümeleme ve sınıflandırma algoritmaları benzeri algoritmalardır.
4. *Yorumlama:* Veri kümesine veri madenciliği algoritmalarının uygulanması sonucunda çıkan bilgilerin yorumlanması işlemidir. Burada elde edilen bilginin özgün, yararlı ve geçerli olup olmadığı değerlendirilmektedir.

5.3. Veri Madenciliğinin Uygulama Alanları

1. *Görüntü İşleme Verileri:* Resimler genellikle sıkıştırılmış olarak ya da ham halleri ile büyük veri tabanlarında saklanırlar. Resimlerin saklanması sayısal ya da metin şeklinde olmaktadır. Bu verilerin işlenmesi için doğru indeksleme, doğru seçimler ve farklılıkların ortaya çıkarılması gerekmektedir.
2. *Sağlık Verileri:* İnsanların yaşam kalitesinin arttırılması ve insan ömrünün uzatılması için sağlık alanında birçok çalışma yapılmaktadır. Bu çalışmalar sonucunda elde edilen veriler büyümekte ve anlamsızlaşmaktadır. Veri madenciliği kullanılarak anlamsız halde bulunan büyük boyuttaki verilerden anlamlı, geçerli ve yeni bilgiler elde edilebilmektedir.
3. *Biyoenformatik:* İnsan vücudu elli bin ile yüz bin arasında gen ve protein moleküllerinden meydana gelmekte ve yaşamını sürdürmektedir. Günümüz dünyasında 6 milyara yakın insan yaşamaktadır. Biyoenformatik, çeşitli genomik veri tabanlarının analiz edilmesi ve değerlendirme işlemlerini yapmaktadır [36]. Bu veri tabanlarının analiz ve değerlendirme işlemlerinde veri madenciliği algoritmaları uygulanmaktadır.
4. *Tıbbi Görüntüleme:* Birçok tıbbi sistemler dijital görüntü ile çalışmaktadır. Bu görüntüler her gün depolanmakta ve işlenmektedir. Bu verilerin işlenmesi tıp çalışanları için oldukça önemlidir. Veri madenciliği kullanılarak veriler otonom şekilde işlenebilmektedir [37].

5. *Bankacılık:* Bankalar gün içerisinde yapılan bütün işlemleri depolamaktadır. Depolanan bu verilerden kredi kartı dolandırıcılıklarının belirlenmesi, kredi kartı harcamalarına göre müşterilerin gruplanması, kredi başvurularının değerlendirilmesi vb. birçok problemde veri madenciliği uygulanmaktadır.
6. *Sigortacılık:* Poliçe başvurularında müşterinin değerlendirilmesi, yapılan dolandırıcılık tespitleri gibi birçok sorunun çözümünde sigorta alanında veri madenciliği kullanılmaktadır.
7. *Pazarlama ve Reklam:* Müşterilerin satın alma güçlerinin belirlenmesi, müşterilerin ilgi alanları potansiyel alıcı oldukları ürünlerin belirlenmesi, mevcut müşterilerin elde tutulması, yeni müşterilerin işletmeye çekilmesi, pazar analizi, müşteri analizi ve ürünün piyasa sürülmesinden önce satış tahminleri yapılması gibi birçok alanda veri madenciliği kullanılmaktadır.
8. *Web Verileri:* Günümüzde belki de hacim ve karmaşıklık yönünden en hızlı artan veri web üzerinde bulunmaktadır. Veri madenciliği bu yüzden web verileri için vazgeçilmez bir çözümdür.

5.4. Veri Madenciliğinde Kullanılan Modeller ve İşlevselliği

Veri madenciliği algoritmaları tahmin etmek veya tanımlama yapmak amacıyla kullanılmaktadır. Veri madenciliği modellerini de bu yönüyle iki ana başlık altında toplayabiliriz. Tahmin etmek amacıyla kullanılan modeller; sonuçları bilinen veriler kullanılarak bir modelin geliştirilmesi ve bu model yardımıyla sonuçları bilinmeyen verilerden yeni sonuçların elde edilmesini sağlamaktadır. Tanımlama yapmak amacıyla kullanılan modeller ise; karar vermeyi sağlama amacıyla eldeki veriler içerisinden örüntünün tanımlanmasını sağlamaktadır.

Yaptıkları işlemlere göre veri madenciliği algoritmaları sınıflandırma ve öngörü, kümeleme analizi, birliktelik analizi, bağlantı analizi, örüntü tanıma, ardışık zaman örüntü analizi, aykırı değer analizi, değişim analizi ve dolandırıcılık tespiti şeklinde sıralanabilir.

5.4.1. Sınıflandırma ve Öngörü

Veri madenciliği yöntemleri içinde en sık uygulanan yöntemlerden birisi sınıflandırmadır. Sınıflandırma, öngörü belirleme yöntemidir. Sınıflandırma ve öngörüm model ya da fonksiyon tanımlamayı veya veri sınıfı ya da kavramlarını tanımlama ya da farklılıklarının belirlenmesini amaçlamaktadır [30].

Sınıflandırma ve öngörüm yöntemleri, banka veya sigortacılıkta dolandırıcılık tespitlerinde, pazarlama ve reklam çalışmalarında, hastalık teşhislerinde ve örüntü tanıma gibi birçok alanda kullanılmaktadır.

5.4.2. Kümeleme

Sınıflandırma ve öngörüm yöntemlerinin aksine kümeleme analizi yapılırken sınıflar önceden belirlenmeden verilerin hangi sınıflarda olacağı verilerin birbirlerine olan yakınlıkları ile belirlenmektedir [30]. Kümeleme analizinde başlangıçta kaç sınıf olduğu bilinmediği için sınıf etiketleri eğitilmiş verilerde gösterilmez. Kümeleme yöntemleri bu sınıf etiketlerini üretmek amacı ile kullanılmaktadır. Sınıflar içerisindeki objelerin benzerlikleri maksimumken sınıflar arası benzerlikleri minimum olacak şekilde sınıflandırılmaktadır. Kümeleme analizi yapıldığında bir sınıftaki obje aynı sınıftaki nesnelerle maksimum benzerlik gösterirken, kendisi haricindeki kümelerdeki nesnelerle benzerliği minimum düzeyde olacaktır.

5.4.3. Birliktelik Analizi

Çeşitli niteliklerden oluşan bir veri dizisi içerisinden belli niteliklerin ilişkilerine dair çeşitli kurallar oluşturulmasına birliktelik analizi, bu kurallara ise birliktelik kuralları denir. Örneğin bir arama motoru yaparken birliktelik analizi kullanırsak; arama yaparken analiz kelimesini yazdığımızda, analiz kelimesinin en çok hangi kelimelerle kullanıldığına bakacak. Birliktelik kuralları kullanarak analiz nedir, analiz merkezi, analiz yayınları gibi en yakın sonuçları bulacak ve muhtemel en yakın seçenek ile tamamlayacaktır. Bu işlemler tamamen otonom şekilde olacak ve elde bulunan veri tabanından bu eşleştirmeyi yapacaktır.

5.4.4. Aykırı Değer Analizi

Veri tabanındaki veriler, veri modeline ya da genel davranışa uyum sağlamayan nesneler içerebilir. Bu nesnelere ayrık veriler denilmektedir. Veri madenciliği yöntemlerinin çoğunluğu bu ayrık verileri gürültü ya da istisna olarak kabul etmektedir. Fakat, dolandırıcılık tespiti gibi bazı uygulamalarda, nadir gerçekleşen olaylar sık tekrar eden olaydan daha ilgi çekici gelebilir. Bu aykırı olayların tespitini yapmaya veri madenciliğinde aykırı değer analizi (outlier analysis) denilmektedir [30].

5.4.5. Evrim Analizi

Evrim analizi, nesnelerin eğilimlerinin ya da düzenliliklerinin zamanla davranışlarının değişikliklerini tanımlama veya modelleme için kullanılmaktadır. Evrim analizi, zamanla bir verinin karekterizasyon, ayırım, birliktelik, sınıflandırma ve kümeleme, birliktelik ve korelasyon analizi, sıralı ya da periyodik olarak desen işleme veya benzerlik tabanlı veri analizi gibi yöntemleri içerebilir.

5.5. KNN Algoritması

K-en yakın komşu algoritması (K-Nearest Neighbor-KNN) ilk olarak 1950'li yılların başında tanımlanmıştır [38]. KNN algoritması ilk başlarda büyük eğitim setlerine verilmiş ve bilgisayarlarda yeterli işlem gücü olmadığı için popülerlik kazanmamıştır. KNN algoritması 1960'lı yıllardan sonra bilgisayarların işlem gücünün artması ile önem kazanmıştır [30].

KNN algoritmasının amacı etkin özellikleri kullanılarak önceden sınıflandırılması yapılmış nesnelerden oluşan eğitim seti vasıtasıyla, yeni gelen nesnenin özellikleri kullanılarak sınıflandırılmasının yapılmasıdır. Bu öngörüm yapılırken yeni gelen nesnenin eğitim setindeki her bir nesneye olan uzaklığı hesaplanır. Bu uzaklıklar arasından *k* tanesi seçilir ve *k* tane seçilen nesnenin sınıflarına göre yeni gelen nesnenin sınıflandırılması yapılmaktadır. *K* değeri rastgele seçilmektedir. *K* değeri 1 olarak alınırsa algoritma yeni gelen verinin en yakın olduğu veri kümesini bularak sınıflandırmayı tek bir küme

üzerinden yapar. *K* değeri daha büyük değerlerde alınırsa algoritma, alınan *k* değerine göre sınıfları seçer ve isteğe göre ortalama bir değer ya da en çok yakın olduğu kümeyi seçer.

KNN algoritması, veri madenciliği algoritmaları arasında en kolay anlaşılan ve uygulanabilen algoritmalar arasında yer almaktadır. Uzaklık hesaplaması temeline dayandığı için KNN algoritmasının sayısal veri içeren eğitim setleri üzerinde uygulanması kategorik veri içeren eğitim setlerine göre uygulanmasından daha kolaydır.

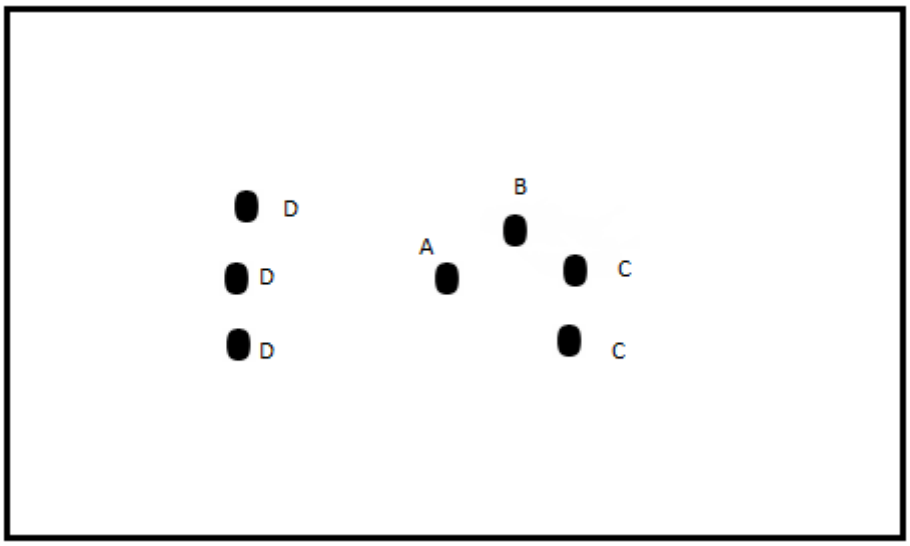
KNN algoritmasının kullanılabilmesi için bütün eğitim setinin, algoritma her çalıştırıldığında bulunması gerekmektedir. Bunun sebebi yeni gelen verinin, eğitim setindeki her bir veri ile olan uzaklığının hesaplanmasıdır. Bunun sonucunda eğitim setinin büyüklüğüne göre sistemde yer kaplaması ve algoritmanın yavaş çalışması gibi dezavantajlar ortaya çıkmaktadır.

KNN algoritmasının çalışma yöntemi;

1. Gelen verinin, eğitim setinde bulunan her bir veriye olan uzaklığını belirle,
2. Belirlenen uzaklıkların sıralamasını yap,
3. Sıralama yapılan uzaklıklar arasında en küçük *k* tane değeri al,
4. Alınan *k* tane değer arasında en çok hangi sınıfta tekrar ediliyorsa gelen veriyi bu sınıfa ata.

5.5.1. K Değerinin Algoritmaya Etkisi

KNN algoritmasında eğitim setinde önceden belirlenen sınıflara göre atama yapıldığı için *k* değeri büyük önem taşımaktadır. Yeni gelen veri eğitim setindeki farklı sınıflara yakın olabilir. Bu veri *k* değeri 1 olarak alınırsa en yakın olduğu sınıfa dahil olacakken, *k* değeri 1'den büyük olarak alındığında yakın olduğu sınıflar arasında en çok olana dahil olacaktır.



Şekil 5.1. KNN algoritmasında *K* değeri değişimi

Eğer örnek verecek olursak Şekil 5.1'de A verisini sınıfı bilinmeyen bir veri olarak ele alındığında, KNN algoritması uygulandığı zaman *k* değerini 1 olarak alınırsa, A verisi B sınıfına dâhil olacaktır. A verisi *k* değerini 3 olarak alınırsa, C sınıfına dâhil olacağı, *k* değeri 6 olarak alındığı zaman ise, D sınıfına dâhil olacağı görülmektedir.

5.5.2. KNN Algoritmasında Verilerin Optimize Edilmesi İçin Kullanılan Algoritmalar

KNN algoritmasının temelinde uzaklık hesaplaması bulunduğu için eğitim setindeki bulunan verilerin sınıflandırmayı etkilememesi için belirli bir aralıkta tutulması gerekmektedir. Verilerin belirli aralıkta tutulması için kullanılan en yaygın veri dönüştürme algoritmaları min-max normalleştirmesi, z-score ve logaritma kullanarak normalleştirme işlemleri uygulanmaktadır.

5.5.2.1. Min-max Normalleştirmesi

Min-max normalleştirilmesi kullanılarak veriler 0 ile 1 arasında ya da -1 ile 1 aralığında normalleştirilebilirler. Min-max normalleştirmesinde veri kümesinde bulunan veriler içerisinde en büyük değer ve en küçük değer bulunur ve daha sonra normalleştirme için aşağıdaki denklemler kullanılmaktadır.

- [0-1] aralığında normalleştirme yapmak için (5.1):

$$x'_i = \frac{x_i - x_{min}}{x_{max} - x_{min}} \quad \textbf{(5.1)}$$

- [-1-1] aralığında normalleştirme yapmak için (5.2):

$$x'_i = \frac{x_i - (\frac{x_{max} + x_{min}}{2})}{\frac{(x_{max} - x_{min})}{2}} \quad \textbf{(5.2)}$$

5.5.2.2. Z-Score Normalleştirmesi

Z-score normalleştirmesinde verilerin ortalaması ve standart sapması hesaplanarak normalleştirme yapılmaktadır. Z-score kullanarak verilerin normalleştirilmesi için denklem (5.3) kullanılmaktadır.

$$x'_i = \frac{x_i - \bar{x}}{\sigma_x} \quad \textbf{(5.3)}$$

Burada σ_x, x değerlerinin standart sapmasını ifade ederken, $\bar{x}$ ise aritmetik ortalamayı ifade etmektedir.

5.5.2.3. Logaritma Kullanarak Normalleştirme

Normalleştirilmesi istenilen verilerin istenilen değere göre logaritması alınarak yapılan normalleştirme tekniğidir [37]. Tablo 5.1'de, verilerin normalleştirilmesi için kullanılan algoritmaların yaş verileri üzerinde kullanımı gösterilmiştir.

Tablo 5.1. Verilerin optimize edilmesi için kullanılan algoritmaların gösterimi

Veriler	Z-Score	[0,1]	[-1,1]	Logaritma 10
27	-1.4976	0	-1	1.4314
31	-1.2053	0.0952	-0.8095	1.4914
41	-0.4748	0.3333	-0.3333	1.6128
69	1.5706	1	1	1.8388
49	0.1096	0.5238	0.0476	1.6902
48	0.0365	0.5000	0	1.6812
62	1.0592	0.8333	0.6667	1.7924
36	-0.8401	0.2143	-0.5714	1.5563
56	0.6209	0.6905	0.3810	1.7482
56	0.6209	0.6905	0.3810	1.7482

5.5.3. KNN Algoritmasında Kullanılan Uzaklık Hesaplama Yöntemleri

KNN algoritması uygulanırken farklı uzaklık hesaplama yöntemleri uygulanmaktadır. Kullanılan uzaklık hesaplama yöntemlerinden bazıları aşağıda verilmiştir.

5.5.3.1. Şehir Mesafe Uzaklığı (Manhattan Uzaklığı)

İsmini Manhattan şehrinden almıştır. Birbirini dik kesen cadde yapılaşmasına sahip olan Manhattan gibi bir şehirde araç sahibinin sadece yatay dikey olarak gidebileceği bu uzaklık hesaplama yönteminde yatay ve dikey uzaklıkların mutlak değer toplamlarına bakılmaktadır. Şehir mesafe uzaklığı, denklem (5.4)'teki eşitlikle hesaplanmaktadır [37].

$$Manhattan\ Uzaklığı_{i,j} = \sum_{i} |x_i - x_j| \qquad (\mathbf{5.4})$$

5.5.3.2. Chebyshev Uzaklığı

Satranç uzaklığı olarak da bilinen Chebyshev uzaklığı, Rus matematikçi Pafnuty Lvovich Chebyshev ismi ile anılmaktadır. Satranç tahtasında rastgele bir yerde duran şahın, gidebileceği yerin uzaklığının hesaplanmasında kullanılmaktadır [37].

Chebyshev uzaklığından faydalanılarak iki vektörün maksimum uzaklıkları hesaplanabilmektedir. Hesaplama işlemi denklem (5.5)'teki bağıntı ile gerçekleşmektedir.

$$Chebyshev\ Uzaklığı(x, y) = max_i(|x_i - y_i|) \quad \textbf{(5.5)}$$

5.5.3.3. Euclidean Uzaklığı

İki nesne arasındaki uzaklığı hesaplamakta kullanılır. Öznitelik sayısını n ve öznitelik değerini *k* göstermek üzere denklem (5.6)'daki eşitlik ile hesaplanmaktadır [37].

$$Euclid\ Uzaklığı_{i,j} = \sqrt{\sum_{k=1}^{n}(x_{ik} - x_{jk})^2} \quad (\mathbf{5.6})$$

5.5.3.4. Minkowski Uzaklığı

Veri kümesindeki değişkenlerin sayısına p dersek, p sayıdaki iki farklı kümede yer alan verilerin birbirine olan uzaklığı (5.7)'deki denklemle hesaplanmaktadır [37].

$$Minkowski\ Uzaklığı_{ij} = \sqrt[m]{\sum_{k=1}^{p}\left|x_{ik} - x_{jk}\right|^2} \quad (\mathbf{5.7})$$

Burada *m* değeri dereceyi gösterirken, *k* değeri öznitelik değerini ve *p* değeri ise öznitelik sayısını göstermektedir.

Minkowski uzaklığı hesaplanırken *m* değeri en fazla 2 olarak alınmaktadır. Bu da euclidean uzaklığını vermektedir. Eğer *m* değeri 1 olarak alınırsa şehir mesafe uzaklığı elde edilecektir.

5.5.4. KNN Algoritmasının Avantajları ve Dezavantajları

KNN algoritması, eğitim setine ihtiyaç duyan bir algoritmadır. Eğitim setinin çok olması durumunda etkili sonuçlar verebilmektedir. KNN algoritması ayrıca eğitim setindeki verilerde bulunan gürültüden çok az etkilenmektedir. KNN algoritmasının dezavantajları arasında; *k* değerinin rastgele belirlenmesi, gelen verinin eğitim setinde bulunan her bir veriye olan uzaklığının hesaplanmasından dolayı işlem yükü oluşması ve

uzaklığa dayalı bir algoritma olmasından ötürü hangi uzaklık hesaplama yönteminin belirlenememesi gibi sorunlar bulunmaktadır [39].

5.6. Bayesyen Algoritması

Bayesyen algoritması KNN algoritmasında olduğu gibi önceden sınıflara ayrılmış eğitim setindeki verileri kullanarak yeni verinin hangi sınıfa ait olacağını bulmaktadır. Bayesyen algoritması sınıflandırma işlemini yaparken istatistiksel yöntemler kullanmaktadır [33]. Bayesyen algoritması eğitim setinde sınıflara ayrılmış olarak bulunan kategorisel verileri kullanarak, yeni gelen verinin hangi sınıfa ait olduğunu bulmaktadır. Bu işlemi yaparken yeni gelen veriyi, eğitim setindeki her bir veriyi kullanarak olasılık çıkarma işlemine tabi tutmaktadır [39].

5.6.1. Bayes Teoremi

Bayes teoremi 18. yüzyılda toplum kurallarına uymayan din adamı Thomas Bayes'in olasılık ve karar hipotezleri üzerine yaptığı çalışmalarla ortaya çıkmıştır [30]. Bayes teoreminin temel hesaplama denklemi (5.8)'de gösterilmiştir.

$$P(h\backslash D) = \frac{P(D\backslash h)P(h)}{P(D)} \text{ [39]} \quad \textbf{(5.8)}$$

P(h|D) ; D olayının meydana geldiği durumda h olayının gerçekleşme olasılığıdır.
P(D|h) ; h olayının meydana geldiği durumda D olayının gerçekleşme olasılığıdır.
P(h) ve P(D) ; h ve D olaylarının önsel olasılıklarını göstermektedir.

5.6.2. Bayes Sınıflandırıcısı

Sade Bayes sınıflandırıcısı (Naive Bayes Classifier) olarak bilinmektedir. X adında hangi sınıfa ait olduğunu bilmediğimiz yeni bir verimiz olsun. X verimizin $\{x_1,x_2,x_3,\ldots,x_n\}$ şeklinde özellikleri olsun. C isimli bir eğitim setimiz ve bu eğitim setinde m adet $\{C_1,C_2,C_3,\ldots,C_n\}$ şeklinde sınıf olduğunu varsayılırsa;

Bayes teoremine göre X olayı meydana geldiğinde C_i olayının gerçekleşme olasılığı denklem (5.9)'daki gibidir [40]:

$$P(C_i \backslash X) = \frac{P(X \backslash C_i)P(C_i)}{P(X)} \quad \textbf{(5.9)}$$

Hesaplama sırasında daha hızlı sonuç almak için, P(X) olasılığı için sadeleştirme yoluna gidilebilir. Eğer sadeleştirme yapmak istersek X_i değerlerinin birbirinden bağımsız olduğunu düşünülerek denklem (5.10)'daki bağıntı kullanabilir.

$$P(C_i \backslash X) = P(X \backslash C_i)P(C_i) \quad \textbf{(5.10)}$$

Eğer bu işlem sırasında, özelliklerin hepsi bağımsız ise denklem (5.11)'in kullanılması gerekmektedir [40].

$$P(X \backslash C_i) = \prod_{k=1}^{n} P(x_k \backslash C_i) = P(x_1 \backslash C_i) \times P(x_2 \backslash C_i) \ldots \times P(x_n \backslash C_i) \quad \textbf{(5.11)}$$

Sınıfını öğrenmek istediğimiz X'i, P(C_i/X)'deki paydalar eşit olduğu için sadece pay değerlerini kullanarak ait olabileceği sınıfa atanabilmektedir. Elde edilen değerler sonucunda en büyük payda seçilerek hangi sınıfa ait olduğu denklem (5.12)'yi kullanarak rahatlıkla saptanabilir [40].

$$\arg max = \{P(X \backslash C_i)P(C_i)\} \quad \textbf{(5.12)}$$

Tablo 5.2. Bayesyen algoritması örneği

Sıra	Güven	Sevgi	İlgi	Durum
1	Çok	Orta	Çok	Ayrılık
2	Orta	Çok	Çok	Evlilik
3	Az	Çok	Çok	Ayrılık
4	Çok	Çok	Çok	Evlilik
5	Orta	Orta	Çok	Evlilik
6	Çok	Orta	Az	Evlilik
7	Az	Az	Az	Ayrılık
8	Çok	Az	Çok	Evlilik

Tablo 5.2 ele alındığında güven değeri çok, sevgi değeri çok ilgi değeri az olan bir bireyin verilerinin hangi sınıfa dâhil olacağını Bayesyen sınıflandırıcısı kullanarak hesaplamak istersek:

1. x_1 güven $=$ çok
2. x_2 sevgi $=$ çok
3. x_3 ilgi $=$ az

Bayesyen olasılıklarını her bir durum için aşağıdaki olasılık tablosunu oluştulur.

Tablo 5.3. Olasılık tablosu

		Ayrılık		**Evlilik**	
		Sayı	**Olasılık**	**Sayı**	**Olasılık**
Güven	**Az**	2	2/3	0	0
	Orta	0	0	2	2/5
	Çok	1	1/3	3	3/5
Sevgi	**Az**	1	1/3	1	1/5
	Orta	1	1/3	2	2/5
	Çok	1	1/3	2	2/5
İlgi	**Az**	1	1/3	1	1/5
	Çok	2	2/3	4	4/5

Tablo 5.3'ten faydalanarak;

Evlilik sınıfındaki her bir veri için koşullu olasılık hesabı;

$$P(x_1 \backslash C_1) = P(Güven = çok \backslash Sınıf = evlilik) = 3/5$$

$$P(x_2 \backslash C_1) = P(Sevgi = çok \backslash Sınıf = evlilik) = 2/5$$

$$P(x_3 \backslash C_1) = P(İlgi = az \backslash Sınıf = evlilik) = 1/5$$

Ayrılık sınıfındaki her bir veri için koşullu olasılık hesabı;

$$P(x_1 \backslash C_2) = P(Güven = çok \backslash Sınıf = ayrılık) = 1/3$$

$$P(x_2 \backslash C_2) = P(Sevgi = çok \backslash Sınıf = ayrılık) = 1/3$$

$$P(x_3 \backslash C_2) = P(İlgi = çok \backslash Sınıf = ayrılık) = 1/3$$

şeklinde hesaplanmaktadır. Denklem (5.11)'deki bağıntı kullanılarak;

Evlilik için olasılığı: (3/5)×(2/5) × (1/5) = 6/125

Ayrılık için olasılığı: (1/3) × (1/3) × (1/3) = 1/27 hesaplanır.

Her bir durumun veri setindeki olma olasılığı hesaba katılırsa, evlilik oranı 5/8, ayrılık oranı 3/8 olarak gelecektir.

Evlilik sınıfında olma olasılığı : $(6/125) \times (5/8) = 0.03$

Ayrılık sınıfında olma olasılığı : $(1/27) \times (3/8) = 0.01$ olarak hesaplanmaktadır.

Bulunan sonuçların ışığında denklem (5.12) uyguladığında 0.03 olasılığı ile evlilik sınıfında yer almaktadır.

5.6.3. Bayesyen Algoritması Sayısal Nitelik Değeri

Bayesyen algoritmasında yaş gibi sayısal değerler varsa bu sayısal veriler algoritmanın çalışma kararlılığını bozacaktır. Sayısal verilerin algoritma için uygun hale getirilmesi gerekmektedir.

Eğitim setinde bulunan sayısal verilerin düzgün dağıldığı varsayılırsa denklem (5.13)'deki standart olasılık fonksiyonu kullanılır [30]. σ standart sapmayı gösterirken, μ_{C_i} aritmetik ortalamayı göstermektedir [40].

$$P(X_k \backslash C_i) = \frac{1}{\sqrt{2\pi\sigma_{C_i}}} e^{\frac{(X_k - \mu_{C_i})^2}{2\pi\sigma^2 C_i}} \quad \textbf{(5.13)}$$

Tablo 5.4. Bayesyen sayısal nitelik değeri örneği

	Yaş	**Güven**	**Sevgi**	**Durum**
1	28	Çok	Orta	Ayrılık
2	22	Orta	Çok	Evlilik
3	18	Az	Çok	Ayrılık
4	24	Çok	Çok	Evlilik
5	22	Orta	Orta	Evlilik
6	32	Çok	Orta	Evlilik
7	17	Az	Az	Ayrılık
8	37	Çok	Az	Evlilik

Tablo 5.4 ele alındığında güven durumu çok, sevgi durumu çok yaşı 24 olan verilerin hangi sınıfa dâhil olacağını Bayesyen sınıflandırıcısı ve sayısal nitelik değeri kullanarak hesaplamak istendiğinde:

Evlilik sınıfı için;

Aritmetik ortalama: 27.4- Standart Sapma: 6.78

Hesaplandıktan sonra denklem (5.13)'den faydalanılarak 1.8561 elde edilir:

Ayrılık sınıfı için;

Aritmetik ortalama: 21- Standart Sapma: 6.08

Hesaplandıktan sonra denklem (5.13)'den faydalanılarak 0.8829 elde edilir:

Tablo (5.3'den faydalanılarak) yararlanılarak koşullu olasılıklar hesaplanırsa: Evlilik sınıfı için koşullu olasılık:

$$P(x_2 \backslash C_1) = P(Güven = çok \backslash Sınıf \ = evlilik) = 3/5$$

$$P(x_3 \backslash C_1) = P(Sevgi = çok \backslash Sınıf = \ evlilik) = 2/5$$

Ayrılık sınıfı için koşullu olasılık:

$$P(x_2 \backslash C_2) = P(Güven = çok \backslash Sınıf \ = ayrılık) = 1/3$$

$$P(x_3 \backslash C_2) = P(Sevgi = çok \backslash Sınıf \ = ayrılık) = 1/3$$

Sonuç olarak denklem (5.11) kullanıldığında;

Evlilik sınıfı için olasılığı: (3/5) × (2/5) × (1,8561) × (5/8) = 0.2784

Ayrılık sınıfı için olasılığı: (1/3) × (1/3) × (0,8829) × (3/8) = 0.0360 olarak hesaplanacaktır. Verilen örnek, hesaplamalar sonucunda denklem (5.12) uygulandığında, 0.2784 olasılık ile evlilik sınıfına dâhil olmaktadır.

6. ÇALIŞMADA KULLANILAN VERİLER

2003 yılında İnsan Genom Projesinin tamamlanmasının ardından, genetik faktörlerin ilaç ve metabolizma üzerindeki etkilerini inceleyen bilim dalı olan farmakogenetik, genetik ve genetik dizilimindeki gelişmelerle hızlı bir ivme kazanmıştır. Bunun sonucunda optimum seviyede ilaç kullanmak 21. yüzyıl için büyük bir önem arz etmektedir.

İnsan vücudu için akılcı ilaç kullanımı oldukça önemlidir. İlaç kullanımı eğer ihtiyaç duyulan düzeyde olmazsa aşağıdaki tehlikeler ortaya çıkabilmektedir [41]:

- Etkisiz ve güvensiz tedavi,
- Hastalığın şiddetlenmesi ve uzaması,
- Hasta için sıkıntı ve zarar,
- Tedavi maliyetini arttırmak.

6.1. İlaç Metabolizması

İlaç metabolizması veya biyotransformasyon, ilaçların aktif kısımlarını değiştirerek suda çözünürlüklerini arttırmak için, ilaç etkisi altına alınmış vücudun ortaya koyduğu kimyasal değişimleri tanımlar. İlaçlar bir kez metabolizmaya dahil olduğunda, boşaltım (salgılama) çok daha kolaylaşır. Metabolizma, 1. faz ve 2. faz olmak üzere iki tip reaksiyonla gerçekleşir. 1. faz reaksiyonları genellikle oksidasyon ve redüksiyon reaksiyonlarını içerirken, 2. faz birleşme ve hidrolizi tanımlar [42].

Biyotransformasyon genellikle karaciğerde olmaktadır ancak böbrek, bağırsak, iskelet kası ve plazma önemli alanlarından biri olabilir. İlaç metabolizması aktiviteleri çoğunlukla endoplazmik retikulum veya hücre sıvısında meydana gelirken biyotransformasyon, plazma zarı ve çekirdek zarında da olabilmektedir [43].

Biyotransformasyon genellikle enzimatik olarak kimyasal değişime uğrar ve bunların en önemlisi Sitokrom P450 (CYP450)'dir [44]. CYP, amino asit homolojilerine göre sınıflandırılırlar. Tam uzunluktaki amino asit dizisinin yüzde 40'ı benzerlik gösterdiği için aynı aileden olarak tanımlanmaktadırlar. CYP ailesi insan için 18 formda bulunmakta ve CYP1, CYP2, CYP3... vb. olarak adlandırılmaktadır. Bu enzimler yazı ve tekrar sayı ile

adlandırılmaktadır. Örneğin CYP1A bir altfamilyasını gösterirken, CYP1A1 ise izoformlarından birini göstermektedir [45].

İlaçların metabolize yeteneği özel enzimlerin aktiviteleri tarafından belirlenir. Genleri, iki adet allelden oluşmaktadır, aileden gelen bu genler genellikle harfler ya da rakamlarla ifade edilmektedir. Mutasyon ya da genlerden bir veya ikisinde polimorfizm olması durumunda enzim aktiviteleri değişebilir ve bireysel olarak ilaç etkileri yavaşlayabilir ya da hızlanabilir [46, 47]. Birçok ilaç dozajı genlere ve tedavi yöntemlerine göre belirlenmiştir. Ancak tedavi indeksi dar olan varfarin gibi ilaçlar için dozaj belirlemesi yapılamamaktadır. Günümüzde hala varfarin dozaj belirleme ihtiyacı devam etmektedir.

6.2. Varfarin

Varfarin bir K vitamini antagonisti olup, atriyal fibrilasyonu (kalp boşluklarından kalp kaidesine yakın olan ikisinin normalden farklı çarpmasını) olan hastalarda felcin önlenmesinde, protez kalp kapakçıkları ve kalp damar tıkanıklığı olan hastaları venöz trombolizmden ve akciğer embolisinden korumak için, ayrıca ortopedi ameliyatına giren ve venöz veya atriyal tromboembolizm geçmişi olan hastalarda akciğer embolisinin oluşumunu önleme yöntemi olarak kabul edilen bir tedavidir [48].

Varfarin, güçlü bir vitamin K epoxide reductase VKOR inhibitörüdür. Varfarin, işlevsel K vitamini-bağımlı pıhtılaşma faktörlerinin oluşum süreci için gereken ulaşılabilir VKH2 miktarını düşürerek VKOR'u önlediğinde, anti-pıhtılaşma meydana gelir [49].

6.3. CYP2C9

CYP2C9 geni 10q24 kromozomu üzerinde yer almaktadır [50]. Bu gen üzerindeki genetik çalışmalar, insanlarda ilaç zehirlenmesine neden olduğu ve önemli ölçüde bu enzim tarafından metabolize edilen ilaçların, ilaç metabolizmasını değiştirdiğini göstermiştir [51].

CYP2C9 polimorfizm göstermektedir ve bugüne kadar farklı katalitik aktiviteye sahip enzimleri kodlayan 3 farklı allel varyantı (CYP2C9*1, CYP2C9*2, CYP2C9*3) iyi katogorize edilmiştir [52].

CYP2C9 'un allel varyantları olan CYP2C9*1, CYP2C9*2 ve CYP2C9*3 varfarin üzerinde etkisi olduğu çeşitli çalışmalarda gösterilmiştir [53].

6.4. CYP4F2

CYP4F2 geni 19p13.12 kromozomu üzerinde yer almaktadır. 3 farklı allel varyantı tespit edilmiştir bunlar CYP4F2*1, CYP4F2*2 ve CYP4F2*3'tür. Bu polimorfizmler içerisinden, CYP4F2*1 enzim aktiviteleri ilişkiliyken CYP4F2*2 ve CYP4F2*3'ün enzim aktiviteleri üzerinde etkisi yoktur [54].

6.5. VKORC1

VKORC1 geni 16p11.2 kromozomu üzerinde yer almaktadır. K vitamini kanın pıhtılaşması için gereklidir ancak enzimler aracılığı ile aktif hale getirilmelidir. VKORC1 geni K vitaminin kanın pıhtılaşmasını sağlayan enzimlerin aktivitelerinden sorumludur [55].

6.6. Genetik Olmayan Faktörler

Genetik faktörlerin yanında genetik olmayan faktörlerinde varfarin dozajının belirlenmesinde etkisi vardır. Genetik olmayan faktörler içerisinde yaş, boy ve kilo yer almaktadır.

7. UZMAN SİSTEMİN GELİŞTİRİLMESİ

Uygulamanın ana amacı kanın pıhtılaşmasını önleyen varfarin isimli ilacın genetik (DNA özellikleri) ve genetik olmayan (yaş, boy ve kilo) etkenler sonucunda alınan sonuç verilerinin işlenmesi ile insan ömrünü uzatmak için en uygun dozajın bulunması hedeflenmiştir. Veri madenciliği algoritmalarından KNN ve Bayesyen sınıflandırıcısı kullanılarak uzman sistem geliştirilmiştir. Uzman sisteminin eğitim için kullanılan veri seti Mahmut ÖZER'in "*The Effect Of Polymorphisms In Cytochrome P450 2C9, Cytochrome P450 4F2, Epoxide Hydrolase 1 And Vitamin K Epoxide Reductase 1 On Warfarin Dose In Turkish Patients*" isimli Yeditepe Üniversitesi'nde 2011 yılında sunduğu yüksek lisans tezinden elde edilmiştir.

Tablo 7.1. Çalışmada kullanılan veriler ve verilerin veri tabanında dağılımı

Özellik	Değer	Görülme Sayısı	Görülme Yüzdesi
CYP2C9	w(0)	70	65.4%
	h(1)	28	26.1%
	m(2)	9	8.4%
VKORC1	w(0)	27	25.2%
	h(1)	55	51.4%
	m(2)	25	23.3%
CYP4F2	w(0)	40	37.3%
	h(1)	49	45.7%
	m(2)	18	16.8%
Yaş	0-19	0	0%
	20-39	19	17.7%
	40-59	50	45.7%
	60-79	38	16.8%
	80-100	0	0%
Kanama Hikâyesi	Yok(0)	69	64.4%
	Var(1)	38	35.5%
Kilo	50-59(kg)	9	8.4%
	60-69(kg)	17	15.8%
	70-79(kg)	42	39.2%
	80-89(kg)	23	21.4%
	90-99(kg)	12	11.2%
	100-109(kg)	3	2.8%
	110-119(kg)	1	0.9%
Boy	150-159(cm)	24	22.4%
	160-169(cm)	46	42.9%
	170-179(cm)	29	27.1%
	180-189(cm)	8	7.4%

Veriler, KNN ve Bayes sınıflandırıcısına uygun hale getirmek için CYP2C9, VKORC1 ve CYP4F2 genetik verileri wild, heterozygote ve mutant olarak bulunan veri kümesinden 0 (wild), 1 (heterozygote), 2 (mutant) olarak değiştirilmiştir. CYP2C9'un allel varyantları olan CYP2C9*2 ve CYP2C9*3'ün ikisinin birden enzim aktivitelerini düşürdüğü için tek veri olarak kabul edilmiştir. Bunun sonucunda veri tabanı için Tablo 7.1 elde edilmiştir.

Uygulama sırasında sonuç değerlerinin net elde edilmesi için, 0 mg ile 2.5 mg (dâhil) arasındaki veriler A sınıfına, 2.5 mg ile 5.0 mg (dâhil) veriler B sınıfına, 5.0 mg ile 7.5 mg (dâhil) arasındaki veriler C sınıfına ve 7.5 mg ile 10 mg (dâhil) arasındaki veriler D sınıfına dâhil edilmiştir. Günlük alınması gereken varfarin dozajının veri tabanında görülme yüzdesi Tablo 7.2'deki gibi elde edilmiştir.

Tablo 7.2. Varfarin dozajı

Aralık	Sınıf Sayısı	Görülme Yüzdesi
0<A≤2,5 (mg)	15	14.0
2,5<B≤5 (mg)	41	38.3
5<C≤7.5 (mg)	43	40.1
7,5<D≤10 (mg)	8	7.4

7.1. KNN Algoritmasının Uygulanması

KNN algoritması uzaklığa dayalı bir algoritma olduğu için ilk olarak verilerin optimize edilmesi gerekmektedir. Yaş, boy ve kilo gibi büyük değerli verilerin algoritmayı olumsuz etkilememesi için z-score normalleştirmesi kullanılarak diğer veriler ile limitleri yaklaştırılmıştır. Büyük değerli veriler içerisinden, yaş verisinin 10 adet örnek veri kümesine, z-score uygulandıktan sonra elde edilen değerler Tablo 5.1'de gösterilmiştir.

KNN algoritmasının ikinci adımı olan uzaklığın ölçülmesi için Öklid uzaklığı kullanılmıştır. Öklid uzaklığı hesaplanarak yeni gelen verinin, veri tabanında bulunan verilerin her birine olan uzaklığı hesaplanmıştır. Tablo 7.3'te bu işlemi yapan kod satırı verilmiştir.

Tablo 7.3. KNN algoritması uzaklık hesaplama

```
uzaklik=0;
c=1;
for i=1:1:boyut
    for j=1:1:sss
        uzaklik=uzaklik+abs((X(i,j)-Y(1,j))^2);
    end
    HSU(c)=sqrt(uzaklik);
    c=c+1;
    uzaklik=0;
end
```

Elde edilen uzaklıkları içeren tek boyutlu dizi, küçükten büyüğe sıralanmıştır. İlk halindeki indisleri tutularak yeni gelen verinin hangi değerlere daha yakın olduğu hesaplanmıştır. Sıralanmış uzaklık değerleri arasından *k* tanesi seçilerek yeni gelen verinin A, B, C ve D sınıflarına olan yakınlığı belirlenmiştir. *K* değerine göre sınıf hesaplamasını yapan kod parçası Tablo 7.4'te gösterilmiştir. *K* değeri uzman sistemin geliştirilmesi sırasında 1 olarak alınmıştır.

Tablo 7.4. KNN algoritmasındaki öklid uzaklık değerlerinin sıralanması ve *k* değeri

```
[YHSU,indx]= sort (HSU);
k=?;
for i=1:1:k
    kTane(i)=text(indx(i));
end
a=0;b=0;c=0;d=0;
for i=1:1:k
    if(kTane(k)=='A')
        a=a+1;
    end
    if(kTane(k)=='B')
        b=b+1;
    end
    if(kTane=='C')
        c=c+1;
    end
    if(kTane=='D')
        d=d+1;
    end
end
```

En yakın komşular arasında hangi sınıfların ne kadar bulunduğu hesaplandıktan sonra, en çok tekrar eden sınıf yeni gelen verinin sınıfı olarak belirlenmiştir. KNN algoritmasının sonucunda yeni gelen verinin hangi sınıfa atanacağını gösteren kod parçası Tablo 7.5'te verilmiştir.

Tablo 7.5. KNN algoritması sınıf belirleme

```
sd= [a,b,c,d];
[M,I] = max(sd(:));
if(I==1)
    sonuc='A';
end

if(I==2)
    sonuc='B';
end
if(I==3)
    sonuc='C';
end
if(I==4)
    sonuc='D';
end
```

KNN algoritmasının çalışmasına ait blok diyagramı Şekil 7.1'de verilmiştir. Ek-1'de KNN algoritmasının Matlab kodları verilmiştir.

Şekil 7.1. KNN algoritması akış diyagramı

7.2. Bayesyen Algoritmasının Kullanımı

Bayesyen algoritması uzman sistemde uygulanırken ilk olarak veri setinde sınıfları bilinen eğitim verilerinin her sınıfta ne kadar yoğunluğa sahip olduğu hesaplanmıştır. Hangi sınıftan kaç tane olduğunu hesaplayan kod parçası Tablo 7.6'da verilmiştir.

Tablo 7.6. Bayesyen algoritması sınıf tekrarı

```
asayisi=0;bsayisi=0;csayisi=0;dsayisi=0;
for i=1:1:satir
    if text(i)=='A'
        asayisi=asayisi+1;
    end
    if text(i)=='B'
        bsayisi=bsayisi+1;
    end
    if text(i)=='C'
        csayisi=csayisi+1;
    end
    if text(i)=='D'
        dsayisi=dsayisi+1;
    end
end
```

Daha sonra bulunan toplamlar veri setindeki toplam sınıf sayısına bölünerek ayrı ayrı A, B, C ve D sınıfının veri seti içerisinde bulunma olasılıkları hesaplanmıştır (Tablo 7.7).

Tablo 7.7. Bayesyen algoritması sınıf yoğunlukları

```
PA=asayisi/satir;
PB=bsayisi/satir;
PC=csayisi/satir;
PD=dsayisi/satir;
```

Yeni gelen verinin CYP2C9, VKORC1, CYP4F2 ve kanama hikâyesi verilerinin her biri için, veri setinde eğitim amaçlı bulunan veriler içerisinde her birinin ayrı ayrı bulunma sıklığı hesaplanmıştır. Örneğin Tablo 7.8'deki kod parçasında CYP2C9 için bulunma sıklığı *a1* olarak alınmış ve yeni gelen verinin *a1* olayının A sınıfında olma olasılığı hesaplanmıştır. Hesaplama işlemini yeni gelen verinin bir tane öz niteliği olan *a1* verisinin A sınıfı içerisinde ne kadar tekrar ettiğini gösteren kod parçası Tablo 7.8'de gösterilmiştir.

Tablo 7.8. Bayesyen algoritması yeni gelen verinin olasılığı

```
for i=1:1:satir
    if (text(i)=='A')
        if verisetim(i,1)==Y(1,1)
            a1=a1+1;
        end
        if verisetim (i,2)==Y(1,2)
            a2=a2+1;
        end
        if verisetim (i,3)==Y(1,3)
            a3=a3+1;
        end
        if verisetim (i,4)==Y(1,4)
            a4=a4+1;
        end

     end
end
```

```
AsinifindaOO=(a1/asayisi);
```

Yaş, boy ve kilo gibi sayısal verilerin Bayesyen algoritmasında sağlıklı çalışabilmesi için bu verilere sayısal nitelik değeri uygulanmıştır. Sayısal nitelik değeri uygulanması sırasında her bir sınıfın verilerinin kendi içerisinde aritmetik ortalaması ve standart sapması hesaplanmış ve (5.13)'teki denklem kullanılarak sayısal nitelik değerleri çıkartılmıştır. Bayesyen algoritmasının sayısal nitelik değerini uygulayan kod parçacığı Tablo 7.9'da gösterilmiştir.

Bulunan tekrarlama olasılıkları, her bir sınıfın kendi içerisinde olasılıkları ve sayısal nitelik değerlerinin çarpımlarının (5.11)'deki denklem ile hesaplanması sonucunda elde edilen değerlerin içerisinden en büyük olanı yeni gelen verinin hangi sınıfa ait olduğunu gösterecektir. Olasılık hesabı Tablo 7.10'da gösterilmiştir.

Tablo 7.9. Bayesyen algoritması sayısal nitelik değeri

```
UzaklikA=0;UzaklikB=0;UzaklikC=0;UzaklikD=0;
for i=1:1:satir
    if text (i)=='A'
        UzaklikA= (abs(Aort-verisetim (i,sutunNumarasi)))^2;
    end
    if text (i)=='B'
        UzaklikB= (abs(Bort-verisetim (i,sutunNumarasi)))^2;
    end
    if text (i)=='C'
        UzaklikC= (abs(Cort-verisetim (i,sutunNumarasi)))^2;
    end
    if text (i)=='D'
        UzaklikD= (abs(Dort-verisetim (i,sutunNumarasi)))^2;
    end
end

SSA= sqrt(UzaklikA/(asayisi-1));
SSB= sqrt(UzaklikB/(bsayisi-1));
SSC= sqrt(UzaklikC/(csayisi-1));
SSD= sqrt(UzaklikD/(dsayisi-1));

PSA= (1/(sqrt(2*pi*SSA)))*((exp(Y(1,sutunNumarasi)-
Aort))^2/(2*SSA^2));
PSB= (1/(sqrt(2*pi*SSB)))*((exp(Y(1,sutunNumarasi)-
Bort))^2/(2*SSB^2));
PSC= (1/(sqrt(2*pi*SSC)))*((exp(Y(1,sutunNumarasi)-
Cort))^2/(2*SSC^2));
PSD= (1/(sqrt(2*pi*SSD)))*((exp(Y(1,sutunNumarasi)-
Dort))^2/(2*SSD^2));

End
```

Tablo 7.10. Bayesyen algoritması olasılıklar

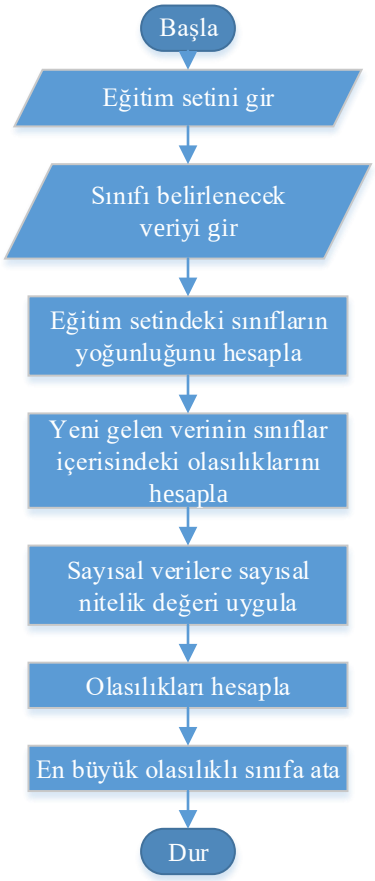
```
OA=(PA*PAO*PSA;
OB=(PB*PBO*PSB);
OC=(PC*PCO*PSC);
OD=(PD*PDO*PSD);
```

Olasılıklar hesaplandıktan sonra olasılıkların sıralanması ve sıralanan bu olasılıklar arasından en büyük olan değerin seçilmesi ile yeni gelen verinin hangi sınıfa atanacağı Tablo 7.11'deki kod satırı ile belirlenmiştir.

Tablo 7.11. Bayesyen algoritması sınıfın belirlenmesi

```
olasiliklar=[OA,OB,OC,OD];
[siraliolasiliklar,indx]= sort (olasiliklar);
if indx(4)==1
    sonuc='A';
end
if indx(4)==2
    sonuc='B';
end
if indx(4)==3
    sonuc='C';
end
if indx(4)==4
    sonuc='D';
end
```

Bayesyen algoritmasının çalışma prensibi Şekil 7.2'de verilmiştir. Ek-2'de Bayesyen algoritmasına ait Matlab kodları verilmiştir.



Şekil 7.2. Bayesyen algoritması akış diyagramı

7.3. Programın Test Edilmesi

Veri setinde bulunan veriler eğitim ve test verisi olarak ikiye ayrılmaktadır. Test verileri 5 kümeden oluşmaktadır. Veri setinde 107 adet veri çeşidi olduğu için test verileri 22, 22, 21, 21 ve 21 adet veri çeşidi olarak ayrılmıştır. Test kümeleri oluşturulurken hafızada indis numaraları tutulmuştur. Bu indis numaraları rastgele oluşturularak her test kümesi için veri setinde geriye kalan veriler eğitim seti olarak kullanılmıştır. Her test verisi kümesindeki elemanlar veri setinden tekrarsız olarak seçilmiştir. Her bir test verisinde bulunan veriler Bayes ve KNN algoritmalarına gönderilmiş ve bir sonuç elde edilmiştir. Elde edilen sonuç, veri setinde bilinen sonuç ile karşılaştırılmıştır. Her bir test kümesi verileri için başarı oranı hesaplanmıştır. Daha sonra hesaplanan bu başarı oranlarının ortalaması alınarak hazırlanan uzman sistemin başarı oranı hesaplanmıştır. Programın test edilmesini gösteren akış diyagramı Şekil 7.3'te verilmiştir. Ek-3'te test işlemini yapan Matlab kodları sunulmuştur.

Küme sayısı *5* olarak alınıp iterasyon sayısı *1000* olarak alındığı zaman Tablo 7.12'deki sonuçlar elde edilmiştir. Ortalama değer iterasyon sayısına göre değişmezken, verilerin az olması ve A ve D sınıflarında eleman sayısının az olmasından dolayı ortalama değer ile maksimum bulunan değer arasında fark oluşmuştur.

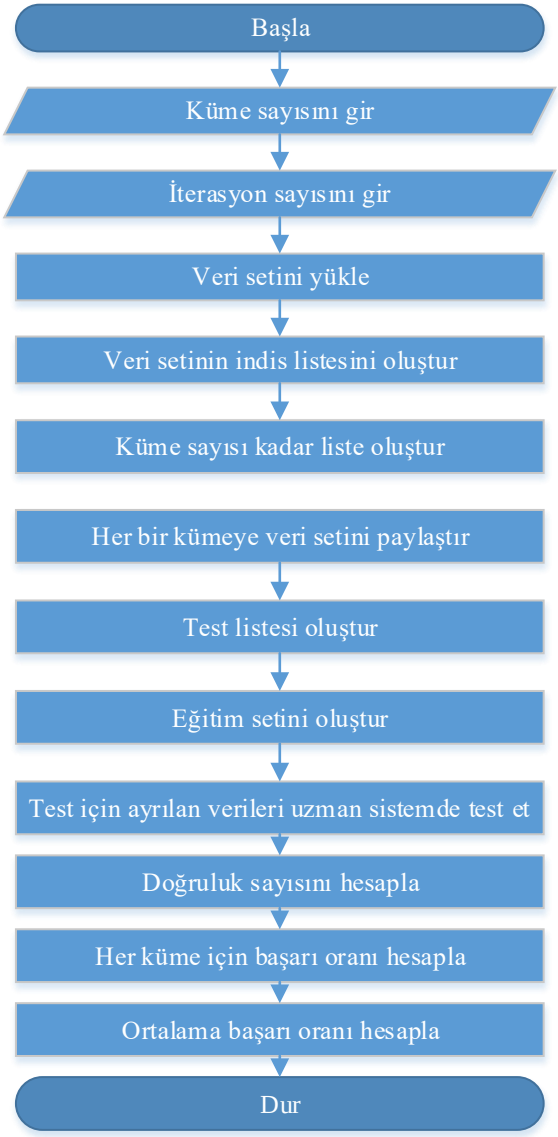
Tablo 7.12. Programın test edilmesi sonucu elde edilen değerler

Algoritma	Ortalama (%)	Maksimum(%)
KNN	37.2545	50.5194
Bayesyen	38.7873	59.0042

Uzman sistemin test edilmesi sırasında maksimum bulunan değerler için 5 farklı veri kümesinde elde edilen oranlar Tablo 7.13'te verilmiştir.

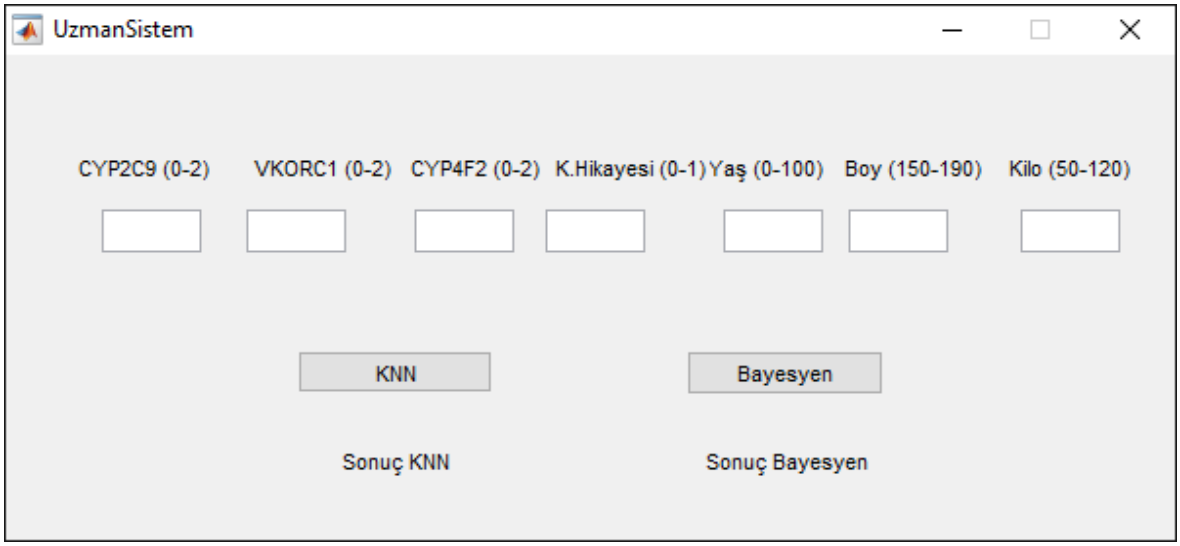
Tablo 7.13. Maksimum elde edilen sonuçlar için her bir küme oranları

	1. Küme	2. Küme	3. Küme	4. Küme	5. Küme	Maksimum
KNN	%40.9090	%54.5454	%38.0952	%52.3809	%66.6666	%50.5194
BAYES	%59.0909	%45.4545	%66.6666	%71.4285	%52.3809	%59.0042



Şekil 7.3. Uzman sistemin test aşaması akış diyagramı

Programın test edilme işlemi bittikten sonra, tasarlanan uzman sistem için ara yüz oluşturulmuştur. Ara yüzde kullanıcı tarafından 7 adet nitelik değeri istenmiştir. Bunlar CYP2C9 (0-2 aralığında), VKORC1 (0-2 aralığında), CYP4F2 (0-2 aralığında), kanama hikayesi (0-1 aralığında), yaş (0-100 aralığında), boy (150-190cm aralığında) ve kilo (50-120kg aralığında)'dur. İstenen bu değerlere göre KNN ve Bayesyen algoritmaları ayrı çalışarak sınıflandırma işlemleri yapılmakta ve kullanıcıya sınıflandırma işlemi sonucu bulunan değer gösterilmektedir. Şekil 7.4'te uzman sistem için tasarlanan ara yüz gösterilmiştir.



Şekil 7.4. Uzman sistemin ara yüz tasarımı

Uzman sistemin 3 farklı hasta için ürettiği sonuçlar Tablo 7.14'te gösterilmiştir. Hasta1 için, CYP2C9 değeri 0, VKORC1 değeri 2, CYP4F2 değeri 0, Kanama Hikâyesi değeri 0, yaş 71, boy 165 ve kilo 75 olarak girildiğinde uzman sistem KNN algoritması için C sınıfına dâhil ederken, Bayesyen algoritması için B sınıfına dâhil etmektedir. Hasta2 için, CYP2C9 değeri 0, VKORC1 değeri 2, CYP4F2 değeri 0, Kanama Hikâyesi değeri 0, yaş 27, boy 175 ve kilo 55 olarak girildiğinde uzman sistem KNN algoritması için A sınıfına dâhil ederken, Bayesyen algoritması için A sınıfına dâhil etmektedir. Hasta3 için, CYP2C9 değeri 1, VKORC1 değeri 1, CYP4F2 değeri 2, Kanama Hikâyesi değeri 0, yaş 55, boy 180 ve kilo 65 olarak girildiğinde uzman sistem KNN algoritması için B sınıfına dâhil ederken, Bayesyen algoritması için B sınıfına dâhil etmektedir.

Tablo 7.14. Çalışmada kullanılan veriler ve verilerin veri tabanında dağılımı

	CYP2C9	VKORC1	CYP4F2	Kanama Hikâyesi	Yaş	Boy	Kilo	KNN	Bayesyen
Hasta1	0	2	0	0	71	165	75	C	B
Hasta2	0	2	0	0	27	175	55	A	A
Hasta3	1	1	2	0	55	180	65	B	B

8. SONUÇLAR

Bilgi, günümüzde gelişen teknoloji ile beraber insanlar arasında hızla yayılmakta ve kontrol edilemez bir biçimde artmaktadır. Hızlı bir şekilde üretilen bu verilerin etkili bir biçimde kullanılabilmesi için anlamlandırılması gerekmektedir. Verilerin anlamlı hale getirilmesi araştırmacılara büyük kolaylık sağlamaktadır.

Genetik alanında araştırma yapan araştırmacılar kullandıkları uygulamalar ve veri boyutları bakımından bilgi teknolojilerine ihtiyaç duyabilmektedir. Genetik alanında yapılan araştırmalar sonucu ortaya çıkan dizilimlerin boyutu gün geçtikçe büyük miktarlarla artmaktadır. Büyük miktardaki bu verilerin anlamlandırılması ve yeni gelen veriler ile kıyaslanması gerekmektedir. Günümüzde biyoenformatik özellikle bu verilerin araştırmalarında vazgeçilmez bir unsur olmuştur.

Bu tez çalışmasında biyoenformatik ve genetik mühendisliği, genom kavramı, uzman sistemler ve veri madenciliği incelenmiştir. Veri madenciliği algoritmalarından KNN ve Bayesyen algoritmaları ayrıntılı bir şekilde açıklanmıştır. Bu algoritmalar kullanılarak varfarin ilacının kullanım miktarı belirlenmeye çalışılarak gereğinden fazla ya da az ilaç kullanımı tasarlanan uzman sistem ile önlenmeye çalışılmıştır. Tasarlanan uzman sistemin daha önceki çalışmalardan en büyük farklılığı, günlük alınması gereken dozaj miktarının miligram olarak değil, tablet olarak hesaplanmasıdır [19]. Sadece Bayesian algoritmasının kullanıldığı önceki çalışmalarda elde edilen başarım oranı %44-56 arasındadır.

Varfarin oranını tespit etmeye yönelik geliştirilen uzman sistemde 107 hasta verisi üzerinden elde edilen CYP2C9, VKORC1 ve CYP4F2 genetik verileri ve genetik olmayan kanama durumu, yaş, boy, ve kilo verileri olmak üzere 7 adet öznitelik değeri kullanılmıştır. Tasarlanan uzman sistem de KNN algoritması için %50.5 ve Bayesyen algoritması için %59.0 başarı oranları elde edilmiştir. Yapılan çalışmada, Bayesyen algoritması, KNN algoritmasına göre kişisel ilaç kullanım dozajının belirlenmesinde daha etkili olduğu ve dozaj miktarı tablet olarak sınıflandırıldığı zaman başarı oranı arttığı tez çalışmasında gözlemlenmiştir.

9. KAYNAKLAR

[1] **Tisdall, J. ,** Beginning Perl for bioinformatics, *O'Reilly Media Inc.*, 2001.

[2] **Frank E., vd.,** Data mining in bioinformatics using weka, *Bioinformatics,* cilt 20, no. 15, pp. 2479-2481, 2004.

[3] **Anderson, N. R., Tarczy-Hornoch, P. ve Bumgarner, R. E.,** On the persistence of supplementary resources in biomedical publications, *BMC bioinformatics* , cilt 7, no. 1, p. 260, 2006.

[4] **Hubbard, T.,** Biological information: making it accessible and integrated (and trying to make sense of it), *Bioinformatics* , cilt 18, no. 2, p. 140, 2002.

[5] **Arndt, T.,** Visual software tools for bioinformatics, *Journal of Visual Languages & Computing*, cilt 19, no. 2, pp. 291-301, 2008.

[6] **Chen, M. ve Hofestädt, R.,** A medical bioinformatics approach for metabolic disorders: Biomedical data prediction, modeling, and systematic analysis, *Journal of biomedical informatics* , cilt 19, no. 2, pp. 147-159, 2006.

[7] **Gabriel, S. B., vd.,** The structure of haplotype blocks in the human genome, *Science,* cilt 296, no. 5576, pp. 2225-2229, 2002.

[8] **Luscombe, N. M., Greenbaum, D. ve Gerstein, M.,** What is bioinformatics? An introduction and overview of the field, *Methods of ınformation In Medicine,* cilt 40, no. 4, pp. 346-358, 2000.

[9] **Akın, A. C., Bürçe, B., Çevirici, B., Şahin, B., Şahin, E. ve Şahin, Y.,** Disiplinler Arası Bir Bilim Dalı: Biyoinformatik, 2010, http://paperzz.com/doc/5078763/p12.-disiplinler-aras%C4%B1-bir-bilim-dal%C4%B1--biyoinformatik, 28 12 2015

[10] **Pauling, L. ve Corey, R.,** Configurations of Polypeptide Chains With Favored Orientations Around Single Bonds: Two New Pleated Sheets, *Proceedings of the National Academy of Sciences of the United States of America* , cilt 37 , no. 11, p. 29–40, 1951.

[11] **G. RC, C. VJ, B. DM, B. B, D. M, D. S, E. B, G. L, G. Y, G. J, H. K, H. T, H. W, I. S, I. R, L. F, L. C, M. M, R. AJ, S. G, S. C, S. G, T. L, Y. JY ve Z. J.,** Bioconductor: open software development for computational biology and bioinformatics, *Genome biology*, cilt 5, no. 10, p. 80, 2004.

[12] **Collins, F. S., Morgan, M. ve Patrinos, A.,** The Human Genome Project: lessons from large-scale biology, *Science*, cilt 300, no. 5617, pp. 286-290, 2003.

[13] wikipedia.org, « DNA,» 2015, http://tr.wikipedia.org/wiki/DNA. ,25 05 2015

[14] **Tuğ, A., Hancı, İ. H. ve Balseven, A.,** İnsan Genom Projesi: Umut mu, Kabus mu?, 2002, http://www.ttb.org.tr/STED/sted0202/genom.pdf. ,07 05 2015

[15] **B. C, M. G, S. B, E. M, S. JI ve B. D.,** A bioinformatics framework for genotype–phenotype correlation in humans with Marfan syndrome caused by FBN1 gene mutations, *Journal of biomedical informatics* , cilt 39, no. 2, pp. 171-183, 2006.

[16] **Dağdelen, A.,** Nükleik Asitler DNA ve RNA Biyoloji Konu Anlatımı, 12 2014., http://www.biyolojihocasi.com/nukleik-asitler-dna-ve-rna-konu-anlatimi, 27 12 2015

[17] **T. JH, W. PS ve H. WL,** THE ORGANIZATION AND DUPLICATION OF CHROMOSOMES AS REVEALED BY AUTORADIOGRAPHIC STUDIES USING TRITIUM-LABELED THYMIDINEE, *Proceedings of the National Academy of Sciences of the United States of America,* cilt 43, no. 1, pp. 122-128, 1957.

[18] **Oflaz, M.**, DNA Replikasyon Mekanizması Nedir?, 02 12 2012, http://www.webmastersitesi.com/webmaster-sozlugu/226523-dna-replikasyon-mekanizmasi-nedir.htm, 13 12 2015

[19] **Oztaner, S., Taskaya, Temizel, T., Erdem, S. ve Ozer, M.,** A Bayesian Estimation Framework for Pharmacogenomics driven Warfarin Dosing: A Comparative Study, *IEEE,* cilt 19, no. 5, pp. 1724 - 1733, 2014.

[20] bioinformaticsweb.net, Bioinformatics and drug discovery, 2005. , http://bioinformaticsweb.net/drugdiscovery.html. ,08 01 2016

[21] **Mitra, S. ve Acharya, T.,** Data mining: multimedia, soft computing, and bioinformatics, John Wiley & Sons, 2005.

[22] **Gupta, S. K.,** A new bioinformatic tool to discover antibiotic resistance genes in bacterial genomes, *Antimicrobial agents and chemotherapy,* cilt 58, no. 1, pp. 212-220, 2014.

[23] **Qin, J. ,vd.,** A human gut microbial gene catalogue established by metagenomic sequencing, *Nature,* cilt 464, no. 7285, pp. 59-65, 2010.

[24] **V. RG, R. SJ ve L. NR,** Cancer gene therapy: hard lessons and new courses, *Gene Ther.,* cilt 7, no. 1, pp. 2-8, 2000.

[25] **Lesk, A.,** Introduction to bioinformatics, İngiltere: Oxford University Press, 2013.

[26] **C. GW, C. TJ, P. JA ve E. E. Jr.**, Biological warfare. A historical perspective, *The Journal of the American Medical Association,* cilt 278, no. 5, pp. 412-417, 1997.

[27] **Kumar, S. ,vd.**, MEGA: a biologist-centric software for evolutionary analysis of DNA and protein sequences, *Briefings in bioinformatics* , cilt 9, no. 4, pp. 299-306, 2008.

[28] **Kumar, S., vd,** MEGA2: molecular evolutionary genetics analysis software,»*Bioinformatics,* cilt 17, no. 12, pp. 1244-1245, 2001.

[29] **Nurminen, J. K., Karonen, O. ve Hätönen, K.,** What makes expert systems survive over 10 years—empirical evaluation of several engineering applications, *Expert Systems with Applications,* cilt 24, no. 2, p. 199–211, 2003.

[30] **Han, J., Kamber, M. ve Pei, J.,** Data Mining: Concepts and Techniques, San Francisco: Elsevier Inc., 2006.

[31] **Wiig, K. M.,** Knowledge Management The Central Management Focus for Intelligent-Acting Organizations, Arlington, Texas: SCHEMA PRESS, 1994.

[32] **Dhaliwal, J. S. ve Benbasat, I.,** The use and effects of knowledgebased system explanations: theoretical foundations and a framework for empirical evaluation, *Information Systems Research,* cilt 7, p. 342–362., 1996.

[33] **Silahtaroğlu, G.,** Veri madenciliği kavram ve algoritmaları, Ankara: Papatya, 2013.

[34] **Klir, G. ve Yuan, B.,** Fuzzy sets and fuzzy logic, cilt 4, New Jersey: Prentice Hall, 1995.

[35] **Fayyad, U.,** Information Visualization in Data Mining and Knowledge Discovery, San Francisco: Morgan Kaufmann Publishers, 2002.

[36] **Durak, B.,** A classification Algorithm Using mahalanobis Distance Clustering of Data With Applications on Biomedical Data Sets, MIDDLE EAST TECHNICAL UNIVERSITY, *Yüksek Lisans Tezi,* Ankara, 2011.

[37] **Akpınar, H.**, Veri Madenciliği Veri Analizi, Ankara: Papatya Yayıncılık, 2014.

[38] **Sillverman, B. W., vd.,** An important contribution to nonparametric discriminant analysis and density estimation, *International Statistical Review,* cilt 57, no. 3, pp. 233-247, 1951.

[39] **Mitchell, T.,** Machine Learning, ABD: McGraw Hill, 1997.

[40] **Özkan, Y.,** Veri Madenciliği Yöntemleri, İstanbul: Papatya, 2013.

[41] **Ambwani ,S. ve Mathur, A. K.,** Rayional Drug Use, *Health Administrator* , cilt 19, no. 1, pp. 5-7, 2006.

[42] **Genç, E. ve Özer, M.,** Importance of individual differences for drug metabolism, *İKU,* cilt 24, pp. 4-6, 2010.

[43] **Brunton, L. L., Lazo, J. S. ve Parker, K. L.,** The Pharmacologic Basis of Therapeutics, New York: McGraw Hil, 2006.

[44] **Ionescu, C. ve Caira, M. R.,** Drug metabolism Current concepts, Springer, 2005.

[45] **Zdanowicz, M. M.,** Concepts in pharmacogenomics, American Society of Health-System Pharmacists®,, Bethesda: ASHP, 2010.

[46] **Guengerich, F. P.,** Catalytic selectivity of human cytochrome P450 enzymes: relevance to drug metabolism and toxicity, *Toxicol. Lett.* , cilt 70, pp. 133-138, 1994.

[47] **Hızel, C.,** Evidence based medicine, predictive and personalized medicine, application of new genetic, *İKU,* cilt 24, pp. 7-18, 2010.

[48] **Özer,M. ,**The Effect Of Polymorphisms In Cytochrome P450 2C9, Cytochrome P450 4F2, Epoxide Hydrolase 1 And Vitamin K Epoxide Reductase 1 On Warfarin Dose In Turkish Patients, Yeditepe University,*Yüksek Lisans Tezi*, İstanbul, 2011.

[49] **Reynolds, K., Valdes, J. R., Hartung, B. R. ve Linder, M. W.,** Individualizing warfarin therapy, *Per. Med.*, cilt 4, no. 1, pp. 11-31, 2007.

[50] **Gray, I. C., Nobile, C., Muresu, R., Ford, S. ve Spurr, N. K.,** A 2.4-megabase physical map spanning the CYP2C gene cluster on chromosome 10q24, *Genomics*, cilt 28, pp. 328-332, 1995.

[51] **Brandolese, R., Scordo, M. G., Spina ,E., Gusella, M. ve Padrini, R.,** Severe phenytoin intoxication in a subject homozygous for CYP2C9*3, *Clin. Pharmacol. Ther.*, cilt 70, pp. 391-394, 2001.

[52] **Miners, J. O. ve Birkett , D. J.,** Cytochrome P4502C9: an enzyme of majör importance in human drug metabolism, *Br. J. Clin. Pharmacol.*, cilt 45, pp. 525-538, 1998.

[53] **Yasar, U., Eliasson, E., Forslund-Bergengren, C., Tybring, G., Gadd, M., Sjöqvist, F. ve Dahl, M. L.**, The role of CYP2C9 genotype in the metabolism of diclofenac in vivo and in vitro, *Eur. J. Clin. Pharmacol*, cilt 57, pp. 729-735, 2001.

[54] **Stec, D. E., Roman, R. J., Flasch, A. ve Rieder, M. J.,** Functional polymorphism in human CYP4F2 decreases 20-HETE production, *Physiol. Genomics*, cilt 30, no. 1, pp. 74-81, 2007.

[55] **S. EA, K. TI, W. HA, A. P, M. L, K. BP, W. P, K. P, D. AK ve K. F.**, The impact of CYP2C9 and VKORC1 genetic polymorphism and patient characteristics upon warfarin dose requirements: proposal for a new dosing regimen, *Epub*, cilt 106, no. 7, pp. 2329-33, 2005.

10. EKLER

Ek-1: KNN arama

```
function [ sonuc ] = knnArama( X,Y, text )
[boyut,sss] = size(X);
uzaklik=0;
c=1;
for i=1:1:boyut
   for j=1:1:sss
      uzaklik=uzaklik+abs((X(i,j)-Y(1,j))^2);
   end
   HSU(c)=sqrt(uzaklik);
   c=c+1;
   uzaklik=0;
end

[YHSU,indx]= sort (HSU);

k=1;

for i=1:1:k
   kTane(i)=text(indx(i));
end
a=0;b=0;c=0;d=0;
for i=1:1:k
   if(kTane(k)=='A')
      a=a+1;
   end
   if(kTane(k)=='B')
      b=b+1;
   end
   if(kTane=='C')
      c=c+1;
   end
   if(kTane=='D')
      d=d+1;
   end
end


sd= [a,b,c,d];
[M,I] = max(sd(:));


if(I==1)
   sonuc='A';
```

```
end

if(I==2)
   sonuc='B';
end
if(I==3)
   sonuc='C';
end
if(I==4)
   sonuc='D';
end
end
```

Ek-2 Bayesyen Arama

```
function [ sonuc ] = bayesArama( X,Y,text )
[satir,sutun]=size(X);
verisetim=X;
asayisi=0;bsayisi=0;csayisi=0;dsayisi=0;
for i=1:1:satir
   if text(i)=='A'
      asayisi=asayisi+1;
   end
   if text(i)=='B'
      bsayisi=bsayisi+1;
   end
   if text(i)=='C'
      csayisi=csayisi+1;
   end
   if text(i)=='D'
      dsayisi=dsayisi+1;
   end
end
PA=asayisi/satir;
PB=bsayisi/satir;
PC=csayisi/satir;
PD=dsayisi/satir;
a1=0;b1=0;c1=0;d1=0;a2=0;b2=0;c2=0;d2=0;a3=0;b3=0;c3=0;d3=0;a4=0;b4=0;c
4=0;d4=0;
for i=1:1:satir
   if (text(i)=='A')
      if verisetim(i,1)==Y(1,1)
         a1=a1+1;
      end
      if verisetim (i,2)==Y(1,2)
         a2=a2+1;
      end
      if verisetim (i,3)==Y(1,3)
         a3=a3+1;
```

```
        end
        if verisetim (i,4)==Y(1,4)
           a4=a4+1;
        end

     end
end
 for i=1:1:satir
    if (text(i)=='B')
        if verisetim(i,1)==Y(1,1)
           b1=b1+1;
        end
        if verisetim (i,2)==Y(1,2)
           b2=b2+1;
        end
        if verisetim (i,3)==Y(1,3)
           b3=b3+1;
        end
        if verisetim (i,4)==Y(1,4)
           b4=b4+1;
        end

     end
 end
 for i=1:1:satir
    if (text(i)=='C')
        if verisetim(i,1)==Y(1,1)
           c1=c1+1;
        end
        if verisetim (i,2)==Y(1,2)
           c2=c2+1;
        end
        if verisetim (i,3)==Y(1,3)
           c3=c3+1;
        end
        if verisetim (i,4)==Y(1,4)
           c4=c4+1;
        end

     end
 end
 for i=1:1:satir
    if (text(i)=='D')
        if verisetim(i,1)==Y(1,1)
           d1=d1+1;
        end
        if verisetim (i,2)==Y(1,2)
           d2=d2+1;
        end
```

```
        if verisetim (i,3)==Y(1,3)
           d3=d3+1;
        end
        if verisetim (i,4)==Y(1,4)
           d4=d4+1;
        end

     end
  end

PAO=(a1/asayisi)*(a2/asayisi)*(a3/asayisi)*(a4/asayisi);
PBO=(b1/bsayisi)*(b2/bsayisi)*(b3/bsayisi)*(b4/bsayisi);
PCO=(c1/csayisi)*(c2/csayisi)*(c3/csayisi)*(c4/csayisi);
PDO=(d1/dsayisi)*(d2/dsayisi)*(d3/dsayisi)*(d4/dsayisi);
[PSA1,PSB1,PSC1,PSD1]=sayisalNitelik(text,verisetim,Y,5);
[PSA2,PSB2,PSC2,PSD2]=sayisalNitelik(text,verisetim,Y,6);
[PSA3,PSB3,PSC3,PSD3]=sayisalNitelik(text,verisetim,Y,7);
OA=(PA*PAO*PSA1*PSA2*PSA3);
OB=(PB*PBO*PSB1*PSB2*PSB3);
OC=(PC*PCO*PSC1*PSC2*PSC3);
OD=(PD*PDO*PSD1*PSD2*PSD3);

olasiliklar=[OA,OB,OC,OD];
[siraliolasiliklar,indx]= sort (olasiliklar);

if indx(4)==1
   sonuc='A';
end

if indx(4)==2
   sonuc='B';
end

if indx(4)==3
   sonuc='C';
end

if indx(4)==4
   sonuc='D';
end
end
```

Bayes Sayısal Nitelik Değerinin Hesaplanması

```
function [PSA,PSB,PSC,PSD] = sayisalNitelik( text,verisetim,Y,sutunNumarasi )
[satir,sutun]=size(verisetim);
asayisi=0;bsayisi=0;csayisi=0;dsayisi=0;
for i=1:1:satir
   if text(i)=='A'
```

```
        asayisi=asayisi+1;
    end
    if text(i)=='B'
        bsayisi=bsayisi+1;
    end
    if text(i)=='C'
        csayisi=csayisi+1;
    end
    if text(i)=='D'
        dsayisi=dsayisi+1;
    end
end
Atoplam=0;Btoplam=0;Ctoplam=0;Dtoplam=0;
for i=1:1:satir
    if (text(i)=='A')
        Atoplam=Atoplam+verisetim(i,sutunNumarasi);
    end
    if (text(i)=='B')
        Btoplam=Btoplam+verisetim(i,sutunNumarasi);
    end
    if (text(i)=='C')
        Ctoplam=Ctoplam+verisetim(i,sutunNumarasi);
    end
    if (text (i)=='D')
        Dtoplam=Dtoplam+verisetim(i,sutunNumarasi);
    end
end
Aort=Atoplam/asayisi;
Bort=Btoplam/bsayisi;
Cort=Ctoplam/csayisi;
Dort=Dtoplam/dsayisi;
UzaklikA=0;UzaklikB=0;UzaklikC=0;UzaklikD=0;
for i=1:1:satir

    if text (i)=='A'
        UzaklikA= (abs(Aort-verisetim (i,sutunNumarasi)))^2;
    end
    if text (i)=='B'
        UzaklikB= (abs(Bort-verisetim (i,sutunNumarasi)))^2;
    end
    if text (i)=='C'
        UzaklikC= (abs(Cort-verisetim (i,sutunNumarasi)))^2;
    end
    if text (i)=='D'
        UzaklikD= (abs(Dort-verisetim (i,sutunNumarasi)))^2;
    end
end

SSA= sqrt(UzaklikA/(asayisi-1));
```

```
SSB= sqrt(UzaklikB/(bsayisi-1));
SSC= sqrt(UzaklikC/(csayisi-1));
SSD= sqrt(UzaklikD/(dsayisi-1));

PSA= (1/(sqrt(2*pi*SSA)))*((exp(Y(1,sutunNumarasi)-Aort))^2/(2*SSA^2));
PSB= (1/(sqrt(2*pi*SSB)))*((exp(Y(1,sutunNumarasi)-Bort))^2/(2*SSB^2));
PSC= (1/(sqrt(2*pi*SSC)))*((exp(Y(1,sutunNumarasi)-Cort))^2/(2*SSC^2));
PSD= (1/(sqrt(2*pi*SSD)))*((exp(Y(1,sutunNumarasi)-Dort))^2/(2*SSD^2));

End
```

Ek3-Programın Test Edilmesi

KNN Test

```
function [ son_oran ] = tezTestFunc( verisetim,text )
elemanSayisi = 107;
kumeSayisi = 5;
sutunSayisi = 7;
liste_eleman = java.util.ArrayList;
for i=1:elemanSayisi
  liste_eleman.add(i);
end

yedek_liste_eleman = liste_eleman.clone();

kumeler = java.util.ArrayList;
egitim = java.util.ArrayList;

for i=1:kumeSayisi-1
   temp1 = java.util.ArrayList;
   kumeler = cat(1,temp1,kumeler);

   temp2 = java.util.ArrayList;
   egitim = cat(1,temp2,egitim);
end


geriSay = elemanSayisi;
say = 0;
for i=1:elemanSayisi
   indis = floor(rand(1)*geriSay)+1;
   deger = liste_eleman.remove(indis-1);
   md = mod(say,kumeSayisi);
   kumeler(md+1).add(deger);
   say = say + 1;
   geriSay = geriSay - 1;
end
for i=1:kumeSayisi
```

```
    egitim(i) = yedek_liste_eleman.clone();
end
for i=1:kumeSayisi
    sayi = kumeler(i).size();
    for j=1:sayi
        eleman = kumeler(i).get(j-1);
        indis = egitim(i).indexOf(eleman);
        egitim(i).remove(indis);
    end
end
toplam = 0;
for i=1:kumeSayisi
sayi1 = kumeler(i).size();
    Y = zeros (sayi1,sutunSayisi);
    T_Test = repmat(char(0),sayi1,1);
    for j=1:sayi1
        eleman = kumeler(i).get(j-1);
        Y(j,:) = verisetim(eleman,:);
        T_Test(j) = text(eleman);
    end
sayi2 = egitim(i).size();
    X = zeros(sayi2,sutunSayisi);
    T = repmat(char(0),sayi2,1);
    for j=1:sayi2
        eleman = egitim(i).get(j-1);
        X(j,:) = verisetim(eleman,:);
        T(j) = text(eleman);
    end

         top = 0;
    for j=1:sayi1

        sonuc = knnArama(X,Y(j,:),T);
        if sonuc == T_Test(j)
            top = top + 1;
        end
    end

    oran = (top / sayi1) * 100;
    toplam = toplam + oran;

end

son_oran = toplam / kumeSayisi;
end
```

Bayesyen Algoritmasının Test Edilmesi

```
function [ son_oran ] = tezTestBayesFunc( verisetim,text )
elemanSayisi = 107;
kumeSayisi = 5;
sutunSayisi = 7;

liste_eleman = java.util.ArrayList;
for i=1:elemanSayisi
  liste_eleman.add(i);
end
yedek_liste_eleman = liste_eleman.clone();
kumeler = java.util.ArrayList;
egitim = java.util.ArrayList;

for i=1:kumeSayisi-1
   temp1 = java.util.ArrayList;
   kumeler = cat(1,temp1,kumeler);

   temp2 = java.util.ArrayList;
   egitim = cat(1,temp2,egitim);
end
 geriSay = elemanSayisi;
say = 0;
for i=1:elemanSayisi
   indis = floor(rand(1)*geriSay)+1;
   deger = liste_eleman.remove(indis-1);
   md = mod(say,kumeSayisi);
   kumeler(md+1).add(deger);
   say = say + 1;
   geriSay = geriSay - 1;
end

for i=1:kumeSayisi
   egitim(i) = yedek_liste_eleman.clone();
end
for i=1:kumeSayisi
   sayi = kumeler(i).size();
   for j=1:sayi
      eleman = kumeler(i).get(j-1);
      indis = egitim(i).indexOf(eleman);
      egitim(i).remove(indis);
   end
end
toplam = 0;
for i=1:kumeSayisi
sayi1 = kumeler(i).size();
   Y = zeros (sayi1,sutunSayisi);
   T_Test = repmat(char(0),sayi1,1);
```

```
    for j=1:sayi1
        eleman = kumeler(i).get(j-1);
        Y(j,:) = verisetim(eleman,:);
        T_Test(j) = text(eleman);
    end

sayi2 = egitim(i).size();
    X = zeros(sayi2,sutunSayisi);
    T = repmat(char(0),sayi2,1);
    for j=1:sayi2
        eleman = egitim(i).get(j-1);
        X(j,:) = verisetim(eleman,:);
        T(j) = text(eleman);
    end
top = 0;
    for j=1:sayi1

        sonuc = bayesArama(X,Y(j,:),T);
        if sonuc == T_Test(j)
            top = top + 1;
        end
    end

    oran = (top / sayi1) * 100;
    toplam = toplam + oran;
end

son_oran = toplam / kumeSayisi;
end
```

ÖZGEÇMİŞ

1988 Bucak doğumlu Osman ALTAY 2005 yılında Bucak Adem Tolunay Fen Lisesi'nden mezun olduktan sonra aynı yıl kazandığı Konya Selçuk Üniversitesi Bilgisayar Sistemleri Öğretmenliği bölümünden 2011 yılında mezun oldu. 2014 yılında Manisa Celal Bayar Üniversitesi Hasan Ferdi Turgutlu Teknoloji Fakültesi Yazılım Mühendisliği Anabilim Dalına YÖK tarafından ÖYP puanı ile araştırma görevlisi olarak atandı. 2014 yılında, Fırat Üniversitesi, Fen Bilimleri Enstitüsü, Yazılım Mühendisliği Anabilim Dalına ÖYP puanı ile lisansüstü eğitim için YÖK tarafından yerleştirildi. 2015 yılı Ocak ayında, Fırat Üniversitesi, Teknoloji Fakültesi, Yazılım Mühendisliği Anabilim Dalında Araştırma Görevlisi olarak göreve başlamış ve halen bu görevine devam etmektedir.