VERİ MADENCİLİĞİ VE MAKİNE ÖĞRENMESİ KULLANILARAK SUÇ ANALİZİ

Merve ORAKCI

YÜKSEK LİSANS TEZİ
ADLİ BİLİŞİM ANABİLİM DALI

GAZİ ÜNİVERSİTESİ
BİLİŞİM ENSTİTÜSÜ

OCAK 2017

MERVE ORAKCI tarafından hazırlanan "VERİ MADENCİLİĞİ VE MAKİNE ÖĞRENMESİ KULLANILARAK SUÇ ANALİZİ" adlı tez çalışması aşağıdaki jüri tarafından OY BİRLİĞİ / OY ÇOKLUĞU ile Gazi Üniversitesi Adli Bilişim Anabilim Dalında YÜKSEK LİSANS TEZİ olarak kabul edilmiştir.

Danışman: Doç. Dr. Bünyamin CİYLAN
Bilgisayar Mühendisliği Anabilim Dalı, Gazi Üniversitesi
Bu tezin, kapsam ve kalite olarak Yüksek Lisans Tezi olduğunu onaylıyorum/onaylamıyorum …….

Başkan: Prof. Dr. Hadi GÖKÇEN
Endüstri Mühendisliği Anabilim Dalı, Gazi Üniversitesi
Bu tezin, kapsam ve kalite olarak Yüksek Lisans Tezi olduğunu onaylıyorum/onaylamıyorum …….

Üye: Yrd. Doç. Dr. Mustafa SERT
Bilgisayar Mühendisliği Anabilim Dalı, Başkent Üniversitesi
Bu tezin, kapsam ve kalite olarak Yüksek Lisans Tezi olduğunu onaylıyorum/onaylamıyorum …….

Tez Savunma Tarihi: 09/01/2017
Jüri tarafından kabul edilen bu tezin Yüksek Lisans Tezi olması için gerekli şartları yerine getirdiğini onaylıyorum.

……
Doç. Dr. Bünyamin CİYLAN
Bilişim Enstitüsü Müdürü

ETİK BEYAN

Gazi Üniversitesi Bilişim Enstitüsü Tez Yazım Kurallarına uygun olarak hazırladığım bu tez çalışmasında;

- Tez içinde sunduğum verileri, bilgileri ve dokümanları akademik ve etik kurallar çerçevesinde elde ettiğimi,
- Tüm bilgi, belge, değerlendirme ve sonuçları bilimsel etik ve ahlak kurallarına uygun olarak sunduğumu,
- Tez çalışmasında yararlandığım eserlerin tümüne uygun atıfta bulunarak kaynak gösterdiğimi,
- Kullanılan verilerde herhangi bir değişiklik yapmadığımı,
- Bu tezde sunduğum çalışmanın özgün olduğunu,

bildirir, aksi bir durumda aleyhime doğabilecek tüm hak kayıplarını kabullendiğimi beyan ederim.

Merve ORAKCI
09/01/2017

VERİ MADENCİLİĞİ VE MAKİNE ÖĞRENMESİ KULLANILARAK SUÇ ANALİZİ

(Yüksek Lisans Tezi)

Merve ORAKCI

GAZİ ÜNİVERSİTESİ

BİLİŞİM ENSTİTÜSÜ

Ocak 2017

ÖZET

İşlenen suçlar; teknolojinin ve uygarlığın gelişimi ile birlikte nitelik olarak değişmiş ve nicelik olarak da artmıştır. Bu doğrultuda suçlar çözülmesi zor ve daha karmaşık yapılara dönüşmüştür. Geçmişte suçları analiz etmek için geleneksel yöntemleri kullanmak yeterli olsa da günümüzde pek mümkün olmamaktadır. Bununla birlikte, farklı yaklaşımların ve yeni teknolojilerin kullanılması zaruri hale gelmiştir. Çalışma alanına uygunlukları ve verimli sonuçlar üretmelerinden dolayı, veri madenciliği ve makine öğrenmesi, suç analizinde kullanılabilen önemli tekniklerdendir. Çalışmada; suçun tanımından ve suç bilimi olan kriminolojinin ilgi alanlarından yola çıkılarak, veri madenciliği ve makine öğrenmesi tekniklerinin suç analizinde nasıl kullanılabileceği açıklanmıştır. İki tekniğin birlikte kullanımı; yapılan analizlerden elde edilen bulgular doğrultusunda değerlendirilmiştir. Bu bağlamda ilk olarak, Federal soruşturma Bürosu tarafından oluşturulan, Ulusal Vaka Tabanlı Raporlama Sistemi kullanılarak; tecavüz, cinayet ve adam kaçırma vakaları üzerinde analizler gerçekleştirilmiştir. Daha sonra; terörist grupların gerçekleştirdiği eylemlerin ayrıntılı bilgisinden oluşan Küresel Terörizm Veritabanı kullanılarak terörist grubu tahmin sistemi geliştirilmiştir.

Bilim Kodu : 902.1.182
Anahtar Kelimeler : Veri Madenciliği, makine öğrenmesi, suç analizi, kriminoloji
Sayfa Adedi : 78
Danışman : Doç. Dr. Bünyamin CİYLAN

CRIME ANALYSIS USING DATA MINING AND MACHINE LEARNING

(M. Sc. Thesis)

Merve ORAKCI

GAZİ UNIVERSITY

INSTITUTE OF INFORMATICS

January 2017

ABSTRACT

With the development of technology and civilization, crime number have increased and its types have changed. In this direction, crime structures have transformed to more complicated and complexed ones. In the past, it was enough to use conventional methods to analyze crimes but nowadays, it is not possible for this. Besides, it is necessary to use different methods and new technologies. Due to the suitability of the study area and producing efficient results, data mining and machine learning are important methods for crime analysis. In this study, based on definition of crime and crimilogy-crime science-, it is explained how to use data mining and machine learning techniques in crime analysis. Usage of these techniques together was evaulated by obtained results. Firstly, using National Incident-Based Reporting System which was developed by Federal Bureau of Investigation, analyses were performed on the cases of rape, murder and kidnapping. Afterwards, using of Global Terrorism Database which has detailed information on the actions taken by terrorist groups, a terrorist group prediction system was developed.

Science Code : 902.1.182
Key Words : Data mining, machine learning, crime analysis, kriminology
Page Number : 78
Supervisor : Assoc. Prof. Dr. Bünyamin CİYLAN

TEŞEKKÜR

Tez çalışmalarımda ve iş hayatımda bana yardımcı olan ve katkı sağlayan değerli danışmanım Doç. Dr. Bünyamin CİYLAN' a, her zaman yanımda olup bana destek olan aileme ve çalışma arkadaşlarıma teşekkür ederim.

İÇİNDEKİLER

Sayfa

Sayfa

ÇİZELGELERİN LİSTESİ

Çizelge **Sayfa**

ŞEKİLLERİN LİSTESİ

Şekil **Sayfa**

SİMGELER VE KISALTMALAR

Bu çalışmada kullanılmış bazı kısaltmalar, açıklamaları ile birlikte aşağıda sunulmuştur.

Kısaltmalar	Açıklama
FBI	Federal Breau of Investigation (Federal Soruşturma Bürosu)
NIBRS	National Incident- Based Reporting System (Ulusal Vaka Tabanlı Raporlama Sistemi)
GTD	Global Terrorism Database (Küresel Terörizm Veritabanı)
DVM	Support Vector Machine (Destek Vektör Makineleri)
NB	Naive Bayes
YSA	Artificial Neural Networks (Yapay Sinir Ağları)
START	National Consortium for the Study of Terrorism and Responses to Terrorism (Terörizm ve Terörle Mücadeleye Yönelik Ulusal Konsorsiyum)
ARFF	Attiribute Relation File Format (Özellik İlişkili Dosya Biçimi)
CSV	Comma Separated Values (Virgülle Ayrılmış Değerler)

1. GİRİŞ

Teknolojinin gelişimi ile birlikte, bilgisayar ve bilgisayar sistemleri insan hayatının vazgeçilmez bir parçası haline gelmiştir. Öyle ki günlük hayatta kullandığımız cep telefonlarından, iş yaşamımızdaki bilgisayarlara; çevre organizasyonlarından, askeri alanda kullanılan bilgisayar sistemlerine kadar her alanda bilişim karşımıza çıkmaktadır. Teknolojinin gelişimi izlendiğinde; önceleri sadece veri transferi yapmak, ya da çok da karmaşık olmayan hesaplamaları yapmak için kullanılan bilgisayar sistemlerinin, zaman içinde evrilerek; büyük miktarda verileri özetleyebilen, mevcut verilerden yeni ve değerli bilgiler elde edebilen sistemler haline dönüştüğü görülmektedir. Diğer bir deyişle günümüz bilgisayarları artık, olaylarla ilgili bilgileri toplayabilmekte, bu olaylarla ilgili kararlar verebilmekte ve olaylar arasındaki ilişkileri çözümleyebilmektedirler. Matematiksel olarak formulasyonu kurulamayacak kadar karmaşık olan ve çözümü mümkün olmayan problemler, sezgisel yöntemlerle bilgisayarlar tarafından çözülebilmektedir. Bilgisayarlara bu yetenekleri kazandıran çalışmalar "yapay zeka" olarak bilinmektedir. Bu terim ilk defa 1950'li yıllarda ortaya atılmış ve 40-50 yıllık bir zaman dilimi içerisinde yoğun ilgi görmüştür. Öyle ki, günümüzde hayatın vazgeçilemez parçası olan sistemlerin doğmasına neden olmuştur [1].

Günümüzde kurumlar büyük boyutlarda veri üretmektedir. Bilişim sistemlerinin kullanımı, verilerin saklanması ve yönetilmesi açısından kurumlara büyük kolaylıklar sağlasa da, verilerin anlamlı bilgiler haline dönüştürülmesi daha çok ön plana çıkmaktadır. İşte bu noktada kurumlar; büyük miktardaki verileri filtreleyerek, anlamlı ve değerli olan bilgileri ortaya çıkarmada zorluk yaşamaktadırlar. Geleneksel istatistik yöntemleri ile büyük boyutlardaki veri kümelerini çözümlemek kolay değildir. Bu nedenle verileri işlemek ve çözümlemek için gelişmiş tekniklere ihtiyaç duyulmuştur. Veri madenciliği teknikleri ve bu teknikleri uygulama aşamasında faydalanılan makine öğrenmesi, bu ihtiyacı karşılamak üzere ortaya çıkmıştır [2].

Suçun varoluşu, insanlığın varoluşuna kadar dayanmaktadır. İlk dönemlerde suçlar, yalın ve basitti. Günümüzde ise teknolojinin ve uygarlıkların gelişimi ile doğru orantılı olarak, suçların türleri ve işleniş biçimleri oldukça artmıştır [3]. Suç ne sistematik ne de tamamen tesadüfi bir olgudur [4]. Suçun bilinmezliği, her zaman insanların ilgisini çekmiş, değişik bilim dallarında çalışanlar bu konularla sürekli olarak ilgilenmişlerdir. Bu çalışmalar genel

olarak; suçun oluşumu, suç ile suçlu arasındaki ilişki, bu ilişkide etkin olan faktörler, suçun önlenmesi, suçu kolaylaştıran ortamlar ve koşullar gibi birçok konu ile ilgili olabilmektedir.

Suç analizi kavramı, suçların oluşmadan önlenmesi, mevcut suçların ve suç eğilimlerinin tespit edilmesi ve bu durumlara karşı tedbirlerin alınmasını içermektedir [5]. Suç analizi ile birlikte; suç ve suçlu arasında ilişki kurulabilir, suçların bölgesel dağılımı hakkında bilgi sahibi olunabilir ve en önemlisi suç işlenmeden önce öngörülerde bulunulabilmektedir. Bu bakımdan suç analizi emniyet ve güvenlik alanında önemli bir yere sahiptir [6].

Veri madenciliği ve makine öğrenmesi yaklaşımlarının suç analizinde kullanımının anlaşılabilmesi ve bu kullanımlara örnek oluşturması açısından, daha önce yapılan çalışmalar aşağıda verilmiştir.

Bruin ve arkadaşları; kümeleme tekniği kullanarak, suçlular ve suç davranışlarını tanımlamak için bütün bireylere ait profilleri ikili olarak karşılaştırmışlardır. Bu karşılaştırmada kullanılan önemli nitelikler; suçun ciddiyeti, suç ortamı, suç sıklığı ve suçun süresidir [7].

Nath ve arkadaşları suç örüntülerini tespit etmek ve suç olaylarının çözüm sürecini hızlandırmak amacıyla veri madenciliğinde kümeleme modelini kullanarak suç analizi yapmışlardır. Çalışmada, bir makine öğrenmesi algoritması olan k-Means kullanılarak suç örüntüsünün tespiti yapılmıştır [8].

Saeed ve arkadaşları; Suç faaliyetlerinin veri kümesi üzerinde, olay sonuçlarını ve özelliklerini tahmin etmek için makine öğrenmesi algoritmaları kullanılmıştır. Yapılan çalışmada kullanılan algoritmalar arasında karşılaştırma yapılmış ve hangi algoritmanın suç analizinde kullanılmasının daha doğru sonuçlar vereceği elde edilen bulgular doğrultusunda tartışılmıştır [9].

Takçı ve Hayta tarafından yapılan çalışmada hırsızlık suç tipi ele alınmış ve bu suçu oluşturan unsurlar analiz edilmiştir. Analiz işlemi veri madenciliği modellerinden birliktelik kuralları ile gerçekleştirilmiştir. Elde edilen kurallar ile hırsızlık suçunu oluşturan nitelikler arasındaki ilişkileri gözlemlemek mümkün hale gelmiştir [6].

Almanie ve arkadaşları; zamansal ve mekânsal suça meyilli bölgeleri tespit etmeye çalışmışlardır. Suça meyilli bölgelerde bilinmeyen desenlerin çıkartılması için apriopri algoritmasını kullanmışlardır. Potansiyel suç tiplerinin tahmin edilebilmesi için ise veri madenciliği tekniklerinden sınıflandırmayı kullanmışlardır. Sınıflandırma tekniğinin uygulanması, karar ağaçları ve Naive Bayes algoritmaları ile gerçekleştirilmiştir. Suçun oluşabileceği tehlikeli bölgelere karşı insanların uyarılmasına olanak sağlaması açısından bu çalışma önem arz etmektedir [10].

Yando ve Olafsson; veri madenciliği tekniklerinden sınıflandırmayı kullanılarak cinayet verileri üzerinde bir analiz işlemi gerçekleştirmişlerdir. Yapılan çalışmada, kurban ve suçlu arasındaki akrabalık ilişkisi tahmin edilmeye çalışılmıştır. Ulusal Vaka Tabanlı Raporlama Sistemi kullanılarak; dört farklı algoritma üzerinde sınıflandırma yapılmıştır. Performansları arttırmak için binary sınıflandırma problemleri oluşturulmuştur. Uygulanan bu yaklaşımda; bir akrabalık ilişkisine karşı diğer tüm akrabalık ilişkileri olmak üzere sınıf sayısı indirgenmiştir. Böylelikle doğruluk oranı iyileştirilerek, daha doğru bir tahminleyici oluşturulmuştur [11].

Asmai ve arkadaşları; veri madenciliği modellerinden birliktelik kurallarını kullanarak özel bir bölge için suç oluşumunu incelemişlerdir. Çalışmada yalnızca coğrafik ve demografik nitelikler kullanılmıştır. Bu veriler Machine Learning Repository'den elde edilmiştir. Gelecekte suçun oluşma olasılığının yüksek olduğu bölgelerin tespit edilebilmesinde kullanılabilecek olması ve bu bölgelerde işlenebilecek suçların engellenebilmesine katkıda bulunabilmesi açısından önem arz etmektedir [12].

Arulanandam ve arkadaşları; makine öğrenmesi algoritmaları kullanarak, online gazete makalelerinden suç bilgileri çıkarma işlemi gerçekleştirmişlerdir. Suç tipi olarak hırsızlık suçunu incelemişlerdir. [13].

Baumgartner ve arkadaşları; olay yerinde gözlemlenen davranışlardan suçlu profilinin çıkarımı için bir bayes ağı modeli geliştirmişlerdir. Burada amaç; cinayet suçuna ait, olay yerinin ve suçlu karakteristiklerinin bilgilerini içeren veri tabanında, bilinmeyen ilişkilerin keşfedilmesidir. Oluşturulan bayes ağı modeli, sonuçlandırılmış suçları içeren veri tabanından, suçlu davranış kalıplarını belirlemeye yöneliktir [14].

Iqbal ve arkadaşları tarafından yapılan çalışmada, suç kategorisi tahmin edilmeye çalışılmıştır. Farklı veri tabanlarının birleştirilmesi ile oluşturulmuş suç verilerini içeren veritabanı üzerinde sınıflandırma algoritmaları uygulanmıştır [15].

Sathyadevan ve arkadaşları ise makine öğrenmesi yaklaşımını kullanarak suç oluşabilecek muhtemel yerleri tahmin edebilmiş ve suça eğilimi olan bölgeleri gösterebilmişlerdir. Bu çalışma suçun oluşmadan önlenebilmesine olanak sağlaması ve bu doğrultuda güvenlik güçlerine fayda sağlaması bakımından önemlidir [16].

Yapılan tez çalışmasında, veri madenciliği ve makine öğrenmesi yaklaşımlarının suç analizindeki farklı problemler üzerinde kullanımı ele alınarak çeşitli analizler yapılmıştır. Bu bağlamda ilk olarak, Federal soruşturma Bürosu tarafından oluşturulan, Ulusal Vaka Tabanlı Raporlama Sistemi kullanılarak; tecavüz, cinayet ve adam kaçırma vakaları üzerinde analizler gerçekleştirilmiştir. Bu analizlerde veri madenciliği tekniklerinden sınıflandırma ve birliktelik kuralları kullanılmıştır. Uygulanan makine öğrenmesi algoritmaları ile elde edilen sonuçlar performans açısından değerlendirilmiştir. Daha sonra; terörist grupların gerçekleştirdiği eylemlerin ayrıntılı bilgisinden oluşan Küresel Terörizm Veritabanı kullanılarak terörist grubu tahmin sistemi geliştirilmiştir. Geliştirilen bu yazılımda, makine öğrenmesi algoritmalarından yapay sinir ağları kullanılmış ve bu algoritma ile eğitilen ağda test işlemi gerçekleştirilmiştir. Bu doğrultuda; elde edilen perfomans çıktıları değerlendirilmiştir. Terörist gurubu tahmin sistemi yazılımının geliştirilmesindeki amaç, eylemi gerçekleştiren terörist grubu yüksek doğrulukta tahmin etmektir.

2. SUÇ BİLİMİ (KRİMİNOLOJİ) VE SUÇ ANALİZİ

Suç, insanların var olduğu ilk çağlardan beri var olan bir kavramdır. Bilinen ilk kasten öldürme suçu, Adem ve Havva'nın ilk çocukları Habil'in, kardeşi Kabil tarafından öldürülmesidir [17]. Suçlar ilk dönemlerde basit şekillerde işlenmekteyken, günümüzde bu durum oldukça farklılık göstermektedir. Teknoloji ve uygarlığın sürekli değişimi ve gelişimi, insan yaşamını her anlamda etkilemesiyle birlikte, etkisini suç olgusunda da göstermiştir. Var olan suç türleri artık basitlikten uzaklaşarak, kompleks yapılara bürünmüştür. Suç kavramı ile ilgili günümüze kadar birçok tanım yapılmış, ancak bunlardan en geçerli olanı; "yasaklanan bir eylem ya da eylemler bütünü veya toplum yasaları ile sınırları çizilen bir yükümlülüğün dışına çıkmaktır" tanımıdır [18].

Suç ve suç ile ilgili kavramlar, her dönemde faklı disiplinlerde çalışanlar tarafından ilgi odağı olmuştur. Bu konu ile ilgili çalışmaların artması, tüm bu çalışmaların tek bir çatı altında toplanmasına sebep olmuş ve böylece kriminoloji bilimi ortaya çıkmıştır. Kriminoloji kısaca, suç işleyen ve suça maruz kalan insanı inceleyen bilim olarak tanımlanabilmektedir [19]. Kriminoloji suç olgusunu incelerken, konuyu sebep sonuç ilişkisi açısından ele almakta ve faili suça iten psikolojik, sosyal ve biyolojik etkenleri, failin mağdura göre statüsü, yakınlığı ve diğer sosyal ilişkileri çerçevesinde incelemektedir. Fail, mağdur ve suç üçlüsünün çok boyutlu olarak ele alınması ve incelenmesi kriminolojinin gelişimine yol açmıştır. Bu doğrultuda kriminoloji, suç ve suça ilişkin konuların bilimsel yöntemlerle incelenmesi olarak tanımlanabilmektedir [3].

Kriminolojinin konusu, sadece kanunlarda suç sayılan fiillerle sınırlı değildir. Buna ek olarak kriminoloji, toplumsal normlardan sapan davranışları da incelemektedir [20]. Suç, kanunlar tarafından açık bir şekilde yasaklanan ve karşılığında ceza öngörülen bir eylem iken, sapma toplumsal normlar çerçevesinde öngörülen kabul edilebilirlik sınırları dışına çıkan her türlü davranıştır [21]. Bu sapıcı davranışlar bazen suç sayılan tipik davranış biçimi ile ilgili olabilir, hatta bir suçun nedeni ya da sonucunu oluşturabilirler. Bu doğrultuda kriminoloji, suç olarak nitelendirilen davranış ortaya çıktıktan sonra bu davranışın oluşmasına neden olan sebepleri araştırır ve buna ilişkin kurallar ortaya koyar. Kriminolojinin ilgi alanına girebilecek diğer bir husus ise hiç bir norma aykırı olmayan davranışların incelenmesidir. Buna örnek olarak, bir öğrencinin derslerinin kötü olması

onun suçluluğu ile yakından ilişkili olabilmektedir. Ya da fazla alkol almak suç değilken, bu kişilerin cinsel suçları, şiddet suçlarını, vandalizm ve trafik suçlarını sıklıkla işledikleri saptanmıştır [3].

Suç analizinde önemli bir konu; suçluların bir profile göre temsil edilmesidir. Bu profil suçlu profili analizi yapılarak çıkarılabilmektedir. Yapılan analiz ile, suçlunun kişilik profiline ulaşılabilmektedir. TREVI; suçlu profili çıkarmayı, olayın özelliklerine göre suçun failinin tanımını oluşturma girişimi, olarak tanımlamıştır. Buradaki varsayım; belli kişilik yapılarının benzer davranış kalıpları sergiler ve bu kalıpların bilinmesi suçun incelenmesinde ve potansiyel şüphelilerin değerlendirilmesinde katkı sağlar, şeklindedir [22].

Belli davranışların suçun işlenmesi sırasında ortamda mevcut olup olmaması, faillerin önceden tahmin edilebilmesinde kilit nokta olabilmektedir. Bu sebeple, ilgili davranışlar özenle incelenmelidir. Yeterli sayıda toplanmış vakalar üzerinde, suçlu profili analizleri yapıldığında, failin suçu işlerken gösterdiği davranışlardan suçluya ait başka özelliklerin çıkarılması da mümkün olabilmektedir. Buna örnek olarak, suçlunun parmak izi konusunda tedbirli davranışı, daha önceden bir mahkumiyetinin olduğunun göstergesi olabilmektedir [19]. Buradan suçlunun işlediği suç sırasında gösterdiği önemli diyebileceğimiz davranışlardan hangi suçu da işleyebileceğinin tahmin edilmesi sonucuna da varılabilmektedir.

Suç ve suç ile ilgili kavramların giderek farklılaşması; analiz ve tespit noktasında farklı yaklaşımların ve yeni tekniklerin kullanılması ihtiyacını doğurmuştur. Bu anlamda veri madenciliği modellerinden kümeleme, sapma tespiti, sınıflandırma ve bağlantı analizi suç analizinde kullanılabilecek önemli yaklaşımlardır. Makine öğrenmesi algoritmaları ise, bu modellerin uygulanma aşamasında kullanılabilmektedir. Bilinen suçların kategorilere ayrılmasında, sınıflandırma modeli uygulanabilecek bir yaklaşımken; bilinmeyen suç türlerinde, kümeleme modeli kullanılabilmektedir. Dolandırıcılık ve ağ saldırıları gibi, normal olandan durumlardan farklı olan durumların tespit edilmesinde, sapma tespiti modeli kullanılabilmektedir. Bağlantı analizi modelinde ise incelenen veri kümesi üzerinde; örüntü tanıma, birlikte gerçekleşen olayların tespit edilmesi veya belli bir sırayla belli olayların gerçekleşmesi gibi bağlantılar ve ilişkiler bulunulmaya çalışılmaktadır [6].

3. VERİ MADENCİLİĞİ VE MAKİNE ÖĞRENMESİ

Bilgisayarlarla üretilen veriler tek başına bir anlam ifade etmemektedir. Bu veriler, ancak belli amaçlar doğrultusunda işlendiği vakit anlam ifade etmeye başlamaktadırlar [23]. Burada önemli olan olgu; geçmişe ait olaylara dair daha gizli bilgilerin keşfedilmesidir. Bu doğrultuda, ileriye dönük öngörüler üretebilmeyi sağlayacak modeller oluşturulmaktadır. Bu modellerle, önceden tedbir almayı sağlayacak bir yönetim anlayışına geçilecek ve olası kayıplar önlenebilecektir [24]. Bu yüzden büyük miktardaki veriyi işleyebilen teknikleri kullanmak önemlidir.

3.1. Veri Madenciliği

Eldeki büyük miktardaki veriyi bilgiye ve anlamlı hale dönüştürme veri madenciliği ile yapılabilmektedir [23]. Veri madenciliğinin geniş bir alanı kapsamasından dolayı, bu konu hakkında birçok tanım bulunmaktadır. Bu tanımlara bakıldığında genel olarak veri madenciliğini açıklarken, veri madenciliğinin farklı boyutlarına değinilmiştir. Bu tanımlardan bazıları aşağıda verilmiştir.

- Veri madenciliği; veri tabanlarında, veri ambarlarında veya diğer veri depolarındaki büyük miktarlardaki veri içindeki ilginç bilgileri keşfetme işidir [25].
- Veri madenciliği; elde bulunan ham verinin tek başına sunamadığı bilgiyi ortaya çıkaran veri analizi sürecidir [26].
- Veri madenciliği, büyük veri yığınları arasından gelecekle ilgili tahminde bulunabilmemizi sağlayabilecek bağlantıların, bilgisayar programı kullanarak aranması işidir [27].
- Veri madenciliği; oldukça tahminci anahtar değişkenlerin binlerce potansiyel değişkenden izole edilmesini sağlama yeteneğidir [28].

Veri madenciliği büyük veri topluluklarından anlamlı ve gelecekteki süreçler için karar vermeye yardımcı olacak nitelikteki bilgilerin ede edilme sürecidir [29]. Yani büyük ölçekli veriler arasından değerli olanı elde etme işidir. Bu sayede veriler arasındaki ilişkileri ortaya koymak ve gerektiğinde ileriye yönelik kestirimlerde bulunmak mümkün olmaktadır [2].

Veri madenciliği sürecinde bu bilgiler ortaya çıkarılırken birçok alandan faydalanılabilmektedir. Bu alanlar; veritabanı teknolojileri, istatistiksel yöntemler, algoritmalar, makine öğrenmesi teknikleri ve yapay zeka olarak sayılabilmektedir [29].

Veri madenciliği büyük miktardaki veriler içerisindeki desenleri keşfeden matematiksel algoritmaları kullanmaktadır. Bu algoritmalar yardımıyla hipotezler oluşturmakta ve sonuçları birleştirip yorumlamak için de insan yeteneğini kullanmaktadır. Bu anlamda veri madenciliği sadece bilim değil bir sanat olarak da yorumlanabilmektedir [30].

3.1.1. Veri madenciliği uygulama alanları

Veri madenciliği bankacılık, pazarlama, sigortacılık, sağlık gibi farklı alanlarda uygulanabilmektedir. Veri madenciliğinin çalışma alanı bulmasında sektör farkı gözetilmemektedir. Ancak geniş veri ambarları oluşturulmasına olanak sağlayan satış, sigortacılık, sağlık gibi alanlarında kullanılması daha yaygın ve daha sağlıklıdır. Veri madenciliğinin kullanım alanlarını incelemek, konuya hakimiyet konusunda destek olacaktır. Aşağıda veri madenciliğinin uygulama alanları verilmektedir [31].

Pazarlama yönetimi

- Müşterilerin satın alma örüntülerinin belirlenmesi,
- Müşterilerin demografik özellikleri arasındaki bağlantıların bulunması,
- Posta kampanyalarında cevap verme oranının artırılması,
- Mevcut müşterilerin elde tutulması, yeni müşterilerin kazanılması,
- Pazar sepeti analizi
- Müşteri ilişkileri yönetimi
- Müşteri değerlendirme
- Satış tahmini
- Çapraz satış.

Risk Yönetimi ve Dolandırıcılık Saptama:

- Kredi kartı dolandırıcılığı
- İnternet işlemleri; e-nakit dolandırıcılığı
- Sigorta dolandırıcılığı
- Kara-para aklama
- Bilgisayar sistemleri ve bilgisayar ağlarına girme
- Telefon dolandırıcılığı
- Üyelik abonelik dolandırıcılığı.

İşaret işleme:

Telefon hatlarındaki parazitlenlenmeden dolayı oluşan kaybı ve bunun sonucunda konuşma sırasında ortaya çıkan gürültüyü yok etme.

Biyoloji(DNA sıra analizi):

Hastalıklara yol açan gen sıralama örneklerinin binlerce gen arasından ortaya çıkarılma işidir. Veri madenciliği yardımıyla geliştirilen sıralama örnek analizi ve benzerlik arama yöntemleri DNA verisi üzerinde analiz yapmayı kolaylaştırmaktadır.

Tıp:

- Var olan bir hastalığın teşhisi
- Bilinmeyen hastalık türlerinin ortaya çıkartılması
- Bir hastalık için önemli sayılabilecek yeni bir bulgunun ortaya çıkarılması.

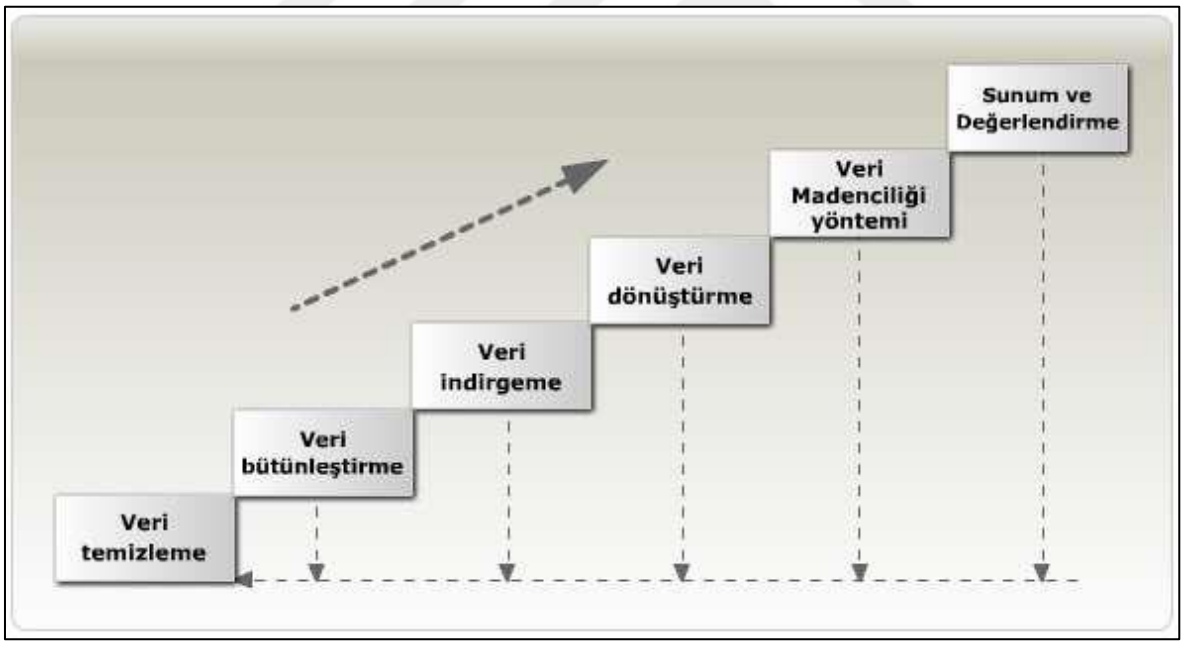
Görüldüğü üzere; veri madenciliği geniş perspektifinden dolayı, birbirinden farklı birçok konuyla ilgilenmektedir. Başka bir deyişle, birbirinden farklı birçok konu veri madenciliği teknikleri ile geliştirilmektedir. Bunu böyle olmasının temel sebebi; veri madenciliği, yapay zeka ve makine öğrenmesi gibi alanların birbirine algoritma ve teknik olarak yaklaşmalarıdır. Aslında veri madenciliği, yapay zeka, makine öğrenmesi adı altında geliştirilen algoritmaları kullanmaktadır. Bu alanlarda çalışanların geliştirdiği algoritmalar da yapay zeka alanına büyük katkı sağlamaktadır [31].

3.1.2. Veri madenciliği süreci

Veri madenciliği denildiğinde çok aşamalı bir süreçten bahsetmek söz konusu olacaktır. Bu süreç aşağıdaki adımları içermektedir [32].

1. Veri temizleme
2. Veri bütünleştirme
3. Veri indirgeme
4. Veri dönüştürme
5. Seçilen algoritmanın veri kümesi üzerine uygulanması
6. Sonuçları sunma ve değerlendirme

Veri madenciliği süreçleri Şekil 3.1'de gösterilmiştir.



Şekil 3.1. Veri madenciliği süreçleri

Veri Temizleme

Veri madenciliği yapılacak ham veri genellikle üzerinde analiz yapılabilecek kadar uygun değildir. Veri kümesi eksik veriler veya analize uygun olmayan tutarsız verileri içerebilmektedir. Bu tutarsız ve hatalı veriler gürültülü olarak değerlendirilmektedir. Böyle durumlarda analizin sağlıklı yapılabilmesi için veri kümesinin söz konusu sorunlardan temizlenmesi gerekecektir. Bu durumda; eksik veri içeren kayıtların silinmesi, değişkenin tüm verileri kullanılarak ortalamasının hesaplanması ve eksik değer yerine bu değerin konması veya eksik değerin tahmin edici algoritmalardan birinin kullanılmasıyla tahmin edilmesi gibi yöntemler kullanılabilmektedir [32].

Veri bütünleştirme

Veri madenciliği süreci büyük veri kümeleri üzerine uygulanmaktadır. Ancak büyük veri kümeleri her zaman direk ulaşılabilir değildir. Bazen veriler dağınık olabilmekte ve bu verilerin birleştirilmesi gerekebilmektedir. Böyle veri tabanlarının birleştirilmesi sürecinde karşılaşılabilecek temel sorun, farklı türdeki verilerin tek türe dönüştürülmesi yani bütünleştirilmesi işlemidir. Aynı türde veriler farklı veri tabanlarında farklı türlerde tutulabilmektedir. Verilerin birleştirilmesi sırasında ise; eğer bütünleştirme yapılmadıysa, aynı veri kümesinde aynı şeyi ifade eden değişkenler sanki farklı değişkenlermiş gibi olabilmekte, bu durum da sağlıklı bir veri madenciliği yapılmasını engellemektedir. Veri madenciliği yapacağımız veri kümesini iyi tanımak bu tür büyük hataların yapılmasını engelleyecektir [2].

Veri indirgeme

Veri madenciliğinde bazen analiz işlemleri uzun sürebilir. Sonuçların değişmemesi kaydıyla veri sayısı ya da değişkenlerin sayısı azaltılabilir. Buna veri madenciliğinde veri indirgeme denmektedir. Ancak burada dikkat edilmesi gereken indirgeme işlemi sırasında analizde önemli sayılabilecek değişkenlerin veya verilerin veri tabanından çıkartılmamasıdır. Bu durumda amaçtan sapma söz konusu olacak ve sonuçlar istenildiği kadar sağlıklı olmayacaktır. Veri indirgeme çeşitli biçimlerde yapılabilmektedir. Bunlar; örnekleme, sıkıştırma, boyut indirgeme veya genelleme gibi biçimlerde yapılabilmektedir [32].

Veri dönüştürme

Veri madenciliği sürecinde her zaman veri, değişkenlerin aynı değerleri kullanılarak analize katılmaz. Bir dönüşüm yöntemi kullanarak değişkenlerin normalleştirilmesi veya standartlaştırılması gerekebilir. Bu durumun oluşmasına sebep olan durumlardan bir kaçı aşağıda verilmiştir:

- Değişkenlerin ortalama ve varyansları birbirlerinden çok farklı olabilir.
- Büyük ortalama ve varyansa sahip değişkenler, diğer değişkenleri analiz sırasında bastırabilir.
- Değişkenler çok küçük veya çok büyük değerlere sahip olabilir.

Yukarıda sayılan sebepler analizlerin sonuçlarını olumsuz etkileyebilmekte ve yanlış değerlendirmelere sebep olabilmektedir. Bu doğrultuda değişkenlerin normalleştirilmesi veya standartlaştırılması analizi daha doğru sonuçlara götürebilmektedir [2].

Min-max normalleştirilmesi

Min-max normalizasyonu verileri 0 ile 1 arasında sayısal değerlere dönüştürmektedir. Her değişken için; en küçük değer (X_{min}) ve en büyük değer (X_{max}) bulunduktan sonra, 3.1 eşitliğindeki formülasyon uygulanmaktadır. Daha sonra her değer yerine normalize edilmiş hali (X_{normal}) yazılmaktadır.

$$X_{normal} = \frac{X - X_{min}}{X_{max} - X_{min}} \tag{3.1}$$

Seçilen algoritmanın veri kümesi üzerine uygulanması

Veri seti ön işlem süreci ile hazır hale geldikten sonra yapılacak olan, probleme uygun yöntemin ve bu doğrultuda algoritmanın seçilerek veri kümesi üzerinde uygulanmasıdır. Veri kümesi üzerine uygulanan her algoritma farklı sonuç üretmektedir. Burada önemli olan probleme en uygun algoritmanın belirlenebilmesidir. Bu algoritmalar; sınıflandırma, kümeleme ve birliktelik kuralları teknikleri altında toplanmaktadır [2].

Sonuçları sunma ve değerlendirme

Veri madenciliği algoritmaları veri kümesi üzerine uygulandıktan sonra yapılacak olan işlem, sonuçların değerlendirilmesi ve bu doğrultuda sunulmasıdır. Değerlendirme ve sunum aşamasında sonuçlar grafiklerle desteklenebilir, kullanılan veri madenciliği platformuna göre elde edilen çıktılar sunum için kullanılabilmektedir. Örneğin hiyerarşik kümeleme yöntemi uygulanmış ise sonuçlar dendrogram grafiği ile sunulmaktadır [33].

3.1.3. Veri madenciliği yöntemleri

Veri madenciliğinde çok sayıda yöntem ve bu doğrultuda bu yöntemleri uygulayabilmek için çok sayıda algoritma geliştirilmiştir. Genel olarak üç aşağıdaki üç modelden bahsetmek doğru olacaktır [2]:

1. Sınıflandırma
2. Kümeleme
3. Birliktelik kuralları

Sınıflandırma

Sınıflandırma modeli veri tabanındaki gizli örüntüleri ortaya çıkarmak için kullanılan bir modeldir. Bu model uygulanırken belli adımlar izlenmektedir. Sınıflandırmaya tabi tutulacak verilerin sınıf etiketleri ve sınıfların sayısı bellidir. Bu tür öğrenmeye danışmanlı (gözetimli) öğrenme denmektedir. Bu yöntemde; verilerin bir kısmı eğitim, bir kısmı ise test için kullanılmaktadır. Veritabanındaki her sınıfın kendi içinde bir deseni vardır. Sınıflandırma modeli bu desenleri keşfederek, bir verinin hangi sınıfa ait olduğunu bulmaya çalışmaktadır. Eğitim kümesi bu desenleri öğrenmek içi kullanılırken, test kümesi ise bu desenlerin ne kadar öğrenildiğini test etmek için kullanılmaktadır. Yapılan test sonucunda doğruluk oranı ne kadar yüksekse model doğrultusunda uygulanan algoritma o kadar başarılı olmuştur. Sınıflandırma modeli en çok bilinen veri madenciliği tekniklerinde birisidir. Resim, örüntü tanıma, hastalık tanıları, dolandırıcılık tespiti, kalite kontrol çalışmaları ve pazarlama konuları sınıflandırma modelinin kullanıldığı alanlardan bazılarıdır. Bu model, tahminleyicidir. Havanın sonraki gün durumunun ne olacağı ya da kutudaki mavi toplarının sayısının tahmin edilmesi, bir sınıflandırma işlemidir. [34].

Kümeleme

Kümeleme verilerin kendi aralarındaki benzerliklerinden hareket edilerek, gruplandırmasıdır. Kümeleme modelinde, sınıflandırmada olduğu gibi sınıf etiketleri belli değildir. Yani bir verinin hangi kümeye ait olduğu belli değildir. Aynı zamanda sınıfların sayısı da belli değildir. Bu tür öğrenmeye danışmansız (gözetimsiz) öğrenme denmektedir. Kümeleme algoritmalarının uygulanmasından sonra sınıflar oluşturulmaktadır. Modelin uygulanmasıyla birbirine en çok benzeyen veriler aynı küme içine konmaktadır. Kümeleme modelindeki amaç, dağınık halde bulunan verileri gruplandırmak ve kullanıcıya anlaşılabilir bir özet veri sunmaktır. Bu modelde; farklı grup içindeki nesneler olabildiğince birbirinden farklı, aynı grup içindeki nesneler ise olabildiğince benzer olacak şekilde oluşturulmaktadır [35].

Birliktelik kuralları

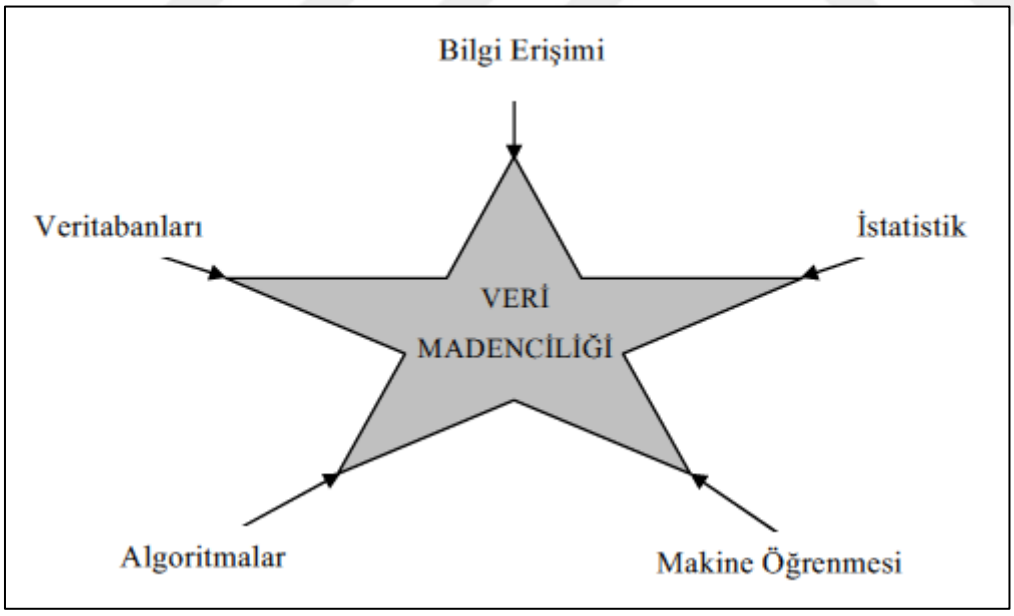
Birlikte olma kurallarını belirli olasılıklarla ortaya koyarak olayların birlikte gerçekleşme durumlarını analiz eden veri madenciliği modeline birliktelik kuralları denmektedir [2]. Birliktelik kuralları yani ilişki analizi, veri kümesindeki bir kaydın diğer kayıtlarla olan bağlantısını açıklayan işlemler dizisidir. Bir kayıt varken bunun yanında başka bir kaydın da var olma olasılığı nedir? Ya da bu iki kayıt varken diğer bir üçüncü veya dördüncü kaydında var olma olasılığı nedir? Şeklindeki sorulara yanıt aramakta ve ortaya çıkardığı bağlantıları birliktelik kuralları olarak tanımlamaktadır [31].

3.2. Makine Öğrenmesi

Makine Öğrenmesini; verilen bir problemi, probleme ait ortamdan edinilen bilgiye göre modelleyen bilgisayar algoritmalarının genel adı olarak tanımlamak mümkündür [36]. Bu yaklaşım ile bilgisayara daha önceki örneklerden edinilmiş tecrübeler öğretilmektedir. Bu doğrultuda makine öğrenmesi, tecrübelerden öğrenme olarak nitelendirilebilmektedir[1]. Makine öğrenmesi yaklaşımlarının bir kısmı tahmin (prediction) ve kestirim (estimation) bir kısmı da sınıflandırma (classification) yapabilme yeteneğine sahiptir [36]. Bu alan, veri madenciliği, istatistik, veri analizi, yapay zeka, bioinformatik ve matematik ile yakından ilişkilidir. Makine öğrenmesinin çeşitli alanlarlarda uygulaması aşağıda verilmektedir [37].

- İstenmeyen e-postaların tanınması (spam detection)
- Dolandırıcılık tespiti(fraud detection)
- Basamak tanıma(digit recognition)
- Konuşma tanıma ve işleme (natural language processing)
- Yüz tanıma (face detection)
- Ürün tavsiye etme (product recommendation)
- Medikal teşhis (medical diagnosis)
- Stok ticareti (stock trading)
- Müşteri bölümleme (customer segmentation)
- Şekil tanıma(shape detection)

Makine öğrenmesi ve veri madenciliği arasında doğrudan bir ilişki bulunmaktadır. Bu ilişki Şekil 3.2' de gösterilmektedir.



Şekil 3.2. Veri madenciliğinin ilişkili olduğu disiplinler [34]

Makine öğrenmesi tekniklerinin büyük veri tabanlarına uygulanması veri madenciliği olarak nitelendirilebilmektedir [1]. Makine öğrenmesi veri madenciliği sürecinde uygulama aşamasında yer almaktadır. Bu aşamada seçilen makine öğrenmesi tekniği veri

seti üzerine uygulanır ve sonuçlar elde edilir. Burada; makine öğrenmesi, öğrenme metotlarını geliştirerek, tahminleri ya da tanımları en iyi şekilde, yüksek performans ile nasıl çıkarılabileceği ile ilgilenirken, veri madenciliği ortaya çıkan bilgi ve bu bilgilerin değerlendirilmesi ile uğraşmaktadır [38].

4. ANALİZLER İÇİN SEÇİLMİŞ ALGORİTMALAR VE VERİ KÜMELERİ

Hedeflenen analizlerin gerçekleştirilebilmesi için; seçilen veri madenciliği tekniğinin uygulanmasında kullanılabilecek, birçok makine öğrenmesi algoritması mevcuttur. Yine üzerinde çalışılacak veri kümesinin kalitesi, elde edilen sonuçların tutarlı ve varılmak istenen hedef doğrultusunda olmasını yüksek oranda etkilemektedir. Bu sebeple seçilen algoritma ve veri kümeleri oldukça önemlidir. Gerçekleştirilen analizler ve terörist grubu tahmin sisteminin oluşturulması için seçilmiş algoritma ve veri kümeleri aşağıda verilmektedir.

4.1. Analizler İçin Seçilmiş Algoritmalar

Gerçekleştirilen analizler ve terörist grubu tahmin sisteminin oluşturulması için, kullanılan teknik ve ulaşılmak istenen hedefler doğrultusunda seçilmiş algoritmalar aşağıda verilmektedir.

4.1.1. Karar ağaçları

Karar ağaçları; danışmanlı öğrenme metotlarından biri olup, karar alma süreçlerinden oluşan bir sınıflandırma algoritmasıdır. Bu algoritma, girdilere göre hedef değişkenin değerini tahmin etmektedir. Karar ağaçlarında, akış şeması şeklinde bir ağaç yapısı mevcuttur. Ağaçta bulunan her düğüm niteliğe dair bir testi temsil etmektedir. Yapraklar ise sınıf etiketlerini göstermektedir ve hedef değişkenin değerinin tahmin edilebilmesi için kararları temsil etmektedirler. Bu kararların oluşturulabilmesi için; eldeki girdilerle kökten yapraklara doğru ilerlemek gerekmektedir. Bu ilerleme sonucunda elde edilen yol, sınıflandırma kurallarını oluşturmaktadır. Oluşturulan bu kurallar, if-else yapıları ile ifade edilebilmektedir [39].

4.1.2. Destek vektör makineleri

Destek vektör makineleri (DVM) sınıflandırma için kullanılan oldukça etkili ve basit yöntemlerden birisidir. Danışmanlı öğrenme tekniklerinden biri olan DVM istatistiksel

öğrenme teorisinden geliştirilmiştir. Eğitim kümesine göre oluşturulmuş model, test kümesi için hedeflenen değerleri tahmin etmektedir. Sınıflandırma yapmak için bir düzlemde bulunan iki grup arasında bir sınır çizilerek iki grubu ayırmak mümkündür. Burada önemli olan; sınırın çizileceği yerin, iki grubun da üyelerine en uzak yer olmasıdır. DVM bu sınırın nasıl çizileceğini belirlemektedir [40].

DVM'de amaç veriyi geniş marjinlerle ayırmaktır [41]. Bu sınıflandırıcılar, marjini maximum yapan optimal ayırıcı aşırıdüzlemi oluşturmaya çalışmaktadır. Margin kavramı, ayırıcı aşırıdüzlemden, en yakın veri noktasına olan minimum uzaklığını ifade etmektedir [42].

DVM, küçük eğitim kümelerinden başarılı genellemeler yapabilmesi ve yüksek boyutlu az miktarda veri için iyi sonuçlar vermesi açısından, güçlü bir sınıflandırma algoritması olarak nitelendirilebilmektedir [43].

4.1.3. Naive bayes

Bayes teoreminin bağımsızlık önermesiyle basitleştirilmiş hali olan Naive bayes (NB), aynı zamanda istatistitiksel bir yöntemdir. NB kullanılarak; sınıfı belli olan veriler yardımıyla, yeni verinin bu sınıflardan herhangi birine girme olasılığı hesaplanmaktadır. NB'de; örüntü tanımada kullanılan her özelliğin istatistiksel olarak bağımsız olması ve aynı derecede önemli olması gerekliliği ortaya konulmaktadır[44].

Farklı sonuçlar üreten sınıflandırma yöntemlerinin hepsi; test veri kümesini, eğitim veri kümesine göre analiz ederek sınıflandırma yapmaktadır [45]. Burada, test verisinin hangi sınıfa gireceği NB yardımı ile belirlenmektedir. Bayes sınıflandırıcı ile; karmaşık yapılı ve yoğun işlemlere gerek kalmadan, büyük veri kümeleri için model oluşturulması kolaydır. Bu doğrultuda; NB modelinin oluşturulması diğere sınıflandırıcılara göre daha kolay olduğu halde, sınıflandırma başarısı bakımından birçok problem tipinde diğer sınıflandırıcıların önüne geçebildiği görülmektedir [46].

4.1.4. Apriopri algoritması

Apriopri algoritması, veri madenciliğinde sık geçen ögelerin keşfedilmesi için kullanılan birliktelik kuralları algoritmasıdır. Sık geçen ögelerin keşfedilmesi, veri tabanının birçok kez taranmasıyla elde edilmektedir. Birliktelik ilişkisi olan ögeler, yapılan bu taramalar sonucunda bulunmaktadır [47].

Sık geçen öğelerin keşfedilmesi için yapılan ilk taramada bir elemanlı minimum destek ölçütünü sağlayan sık geçen öğe kümeleri bulunmaktadır. Yapılan her taramada bir önceki taramada bulunan sık geçen öğe kümeleri, aday kümeler adı verilen bir sonraki sık geçen öğe kümelerini türetmek için kullanılmaktadır. Bu iterasyon yeni bir sık geçen öğe kümesi bulunmayana dek devam etmektedir [32]. Apriopri algoritmasının adımları aşağıda belirtilen aşamalara sahiptir [2]:

1. İlk adım, destek ve güven değerlerini karşılaştırmak için eşik değerlerin belirlenmesidir.
2. Veritabanı taranır ve bir elemanlı minimum destek değerini sağlayan sık geçen öge kümeleri bulunur. Her bir nesne için tekrar sayıları, yani destek sayıları hesaplanır. Elde edilen bu destek sayıları eşik destek sayısı ile karşılaştırılır. Eşik destek sayısından küçük değerlere sahip olan satırlar bu adımda elenir. Koşula uygun kayıtlar, bir elemanlı minimum destek değerini sağlayan sık geçen öge kümelerini oluşturmaktadır.
3. Belirlenen bir elemanlı sık geçen ögeler ikişerli gruplandırılarak bu grupların tekrar sayıları, yani destek sayları elde edilir. Elde edilen bu sayılar eşik destek sayıları ile karşılaştırılır. Eşik değerden küçük satırlar bu adımda elenir. Koşula uygun kayıtlar iki elemanlı minimum destek değerini sağlayan sık geçen öge kümeleridir.
4. Üçerli, dörderli vb. gruplandırmalar yapılır. Elde edilen destek değerleri eşik değerlerle karşılaştırılır. Yeni bir minimum destek değerini sağlayan sık geçen öge kümesi bulunana kadar bu işlemlere devam edilir.
5. Kural destek değerine bakılarak birliktelik kuralları oluşturulur. Oluşturulan kuralların her biriyle ilgili güven değerleri hesaplanır.

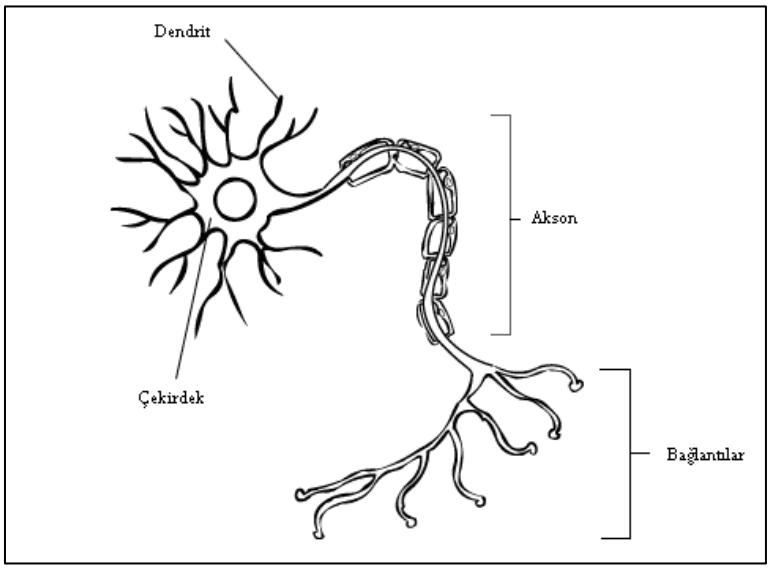
Agrawal ve Srikant tarafından geliştirilen apriopri algoritması 20. VLDB (Very Large Database Endowment) konferansında sunulmuştur. Bildiride sunulan algoritmanın kaba kodu Şekil 4.1' de verilmiştir [48].

```
L1= {sık geçen 1-öğe kümesi};
for (k=2; Lk-1≠ Ø; k++) do begin
    Ck=apriori-gen (Lk-1); // Yeni adaylar
    forall transactions-hareketler t ∈ D do begin
            Ct = subset (Ck, t); // Adaylar t içindedir
            forall candidates – adaylar c ∈ Ct do
                    c.count++;
    end
    Lk = {c ∈ Ck | c.count ≥ minsup}
end
Answer = ∪k Lk;
```

Şekil 4.1. Apriopri Algoritması

4.1.5. Yapay sinir ağları

Yapay sinir ağları (YSA), insan beyninin fonksiyonlarını kullanarak olayları öğrenebilen, çevreden gelen etkilere karşı tepki üretebilen bilgisayar sistemleridir. Tasarlanırken insan beynindeki biyolojik sinir ağlarından esinlenilmiştir [49]. İnsan beynindeki biyolojik sinir hücresi Şekil 4.2'de verilmiştir.



Şekil 4.2. Biyolojik sinir hücresi (nöron)

Öğrenme, ilişkilendirme, sınıflandırma, genelleme, özellik belirleme ve optimizasyonda kullanılabilecek başarılı bir yöntemdir. Eski bilgileri kullanarak oluşturdukları deneyimlerle, gelecekte benzer konularda benzer kararlar verebilmektedirler[50].

Bilgisayarda insan beyni davranışlarını taklit etmek amacıyla tasarlanan [51] yapay sinir ağları, öğrenme sürecinin matematiksel olarak modellenmesi uğraşı sonucu ortaya çıkmıştır. Bu alanda yapılan çalışmalar; beyni oluşturan biyolojik üniteler olan nöronların modellenmesi ve bilgisayar sistemlerinde uygulanması ile başlamış, bilgisayar sistemlerinin gelişimiyle de paralel olarak farklı alanlarda kullanılır hale gelmiştir [52].

YSA'da ki ağırlıklar, eğitim ve test kümeleri kullanılarak eğitilmektedirler. Bu doğrultuda eğitilen YSA yeni bir girdi ile yeni bir çıktı tahmin edebilecek yeterliliğe gelmiş olmaktadır [53].

YSA içerdiği nöronların birbirine bağlanış şekillerine göre ikiye ayrılmaktadır [52]:

- İleri beslemeli ağlar
- Geri beslemeli ağlar

İleri beslemeli ağlarda; bir katmandan sadece kendinden sonraki katmanlara bağ bulunmaktadır, YSA' ya gelen bilgiler önce giriş katmanına, oradan sırasıyla ara katmanlara ve en son çıkış katmanına işlenerek gelir. Çıkış katmanından da dış dünyaya çıkar. Geri beslemeli ağlarda ise ileri beslemeli ağlardan farklı bir durum vardır. Burada bir hücrenin çıktısı kendinden önceki katmanda veya kendi katmanında bulunan herhangi bir hücreye girdi olarak bağlanabilmektedir [52].

Yapay sinir ağında öğrenme işleminin gerçekleştirilmesinde birden fazla yöntem kullanılmaktadır. Bu yöntemler şu şekildedir [54]:

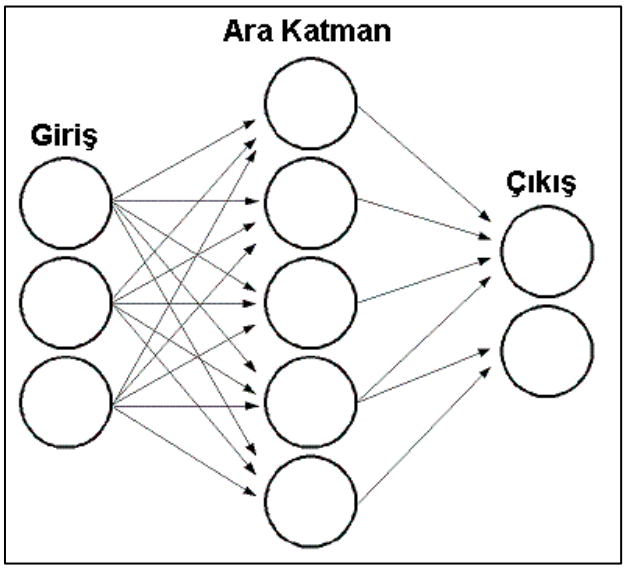
- Danışmanlı öğrenme
- Danışmansız öğrenme
- Destekleyici öğrenme

Danışmanlı Öğrenmede YSA, kullanılmaya başlanmadan önce eğitilmektedir. Eğitim sırasında, hem girdiler hem de girdiler için istenen çıktılar sisteme verilmektedir. Daha sonra, istenen çıktı ile ağın ürettiği çıktı karşılaştırılır ve hata oranı hesaplanır. Bu hesaplamalara göre, ağırlıklar güncellenmektedir. Danışmansız Öğrenmede, sistemin öğrenmesine yardımcı olan herhangi bir danışman yoktur. Sisteme sadece girdiler verilmekte ve sistemin örneklerdeki parametreler arasındaki ilişkileri kendi kendine öğrenmesi beklenmektedir. Pekiştirmeli Öğrenmede ise, girişler ağa uygulanmakta ve sonucun danışman tarafından değerlendirilmesi istenmektedir. Ödüllendirme ve cezalandırma yöntemiyle ağın ağırlıkları güncellenmektedir [54].

Bir yapay sinir ağı, yapay sinir hücrelerinin birbirine bağlanması ile oluşmaktadır. Sinir hücrelerinin bir araya gelmesi rastgele olmamaktadır. Hücreler genel olarak 3 katman halinde ve her katman içinde paralel olarak bir araya gelerek ağı oluşturmaktadırlar. Bu katmanlar aşağıda verilmektedir [1]:

- Girdi katmanı
- Ara katmanlar (hidden layer)
- Çıktı katmanı

Bir yapay sinir örneği Şekil 4.3'de verilmiştir.



Şekil 4.3. Bir yapay sinir örneği

Giriş katmanında girdiler herhangi bir işleme uğramadan ara katmanlara iletilmektedirler. Ara katman sayısı ağdan ağa değişmektedir. Bazı ağlarda ara katman yokken bazılarında birden fazla bulunmaktadır. Ara katmanlardaki nöron sayıları hem giriş-çıkış sayılarından, hem de kendi aralarındaki hücre sayılarından bağımsızdır. Ara katman ve ara katmanlarda bulunan nöron sayıları problemden probleme değişiklik göstermektedir. Ara katmandan çıkış katmanına aktarılan bilgiler, burada da işlenerek dış dünyaya aktarılır. Eğer ağ geri beslemeli ise elde edilen çıktı, ağın yeni ağırlık değerlerinin hesaplanmasında kullanılmaktadır [52].

4.2. Analizler için seçilmiş veri kümeleri

Gerçekleştirilen analizler ve terörist grubu tahmin sisteminin oluşturulması için, seçilmiş veri kümeleri aşağıda verilmektedir.

4.2.1. Ulusal vaka tabanlı raporlama sistemi

Ulusal Vaka Tabanlı Raporlama Sistemi (NIBRS) kanun uygulayıcılar tarafından, suç vakalarını detaylı şeklide kaydetmek için kullanılmaktadır. Bu veriler suç, kurban ve tutukluların detaylı bilgisinden oluşmaktadır. Böyle olay verileri yasa koyucular veya kanun uygulayıcılar için bir bakış açısı sunarak, gizli kalmış desenleri keşfetmek için veri madenciliğinin kullanımına olanak sağlamaktadır [11].

NIBRS, Federal Soruşturma Bürosu (FBI) tarafından yönetilen Düzgün Suç Raporlama Program'ının bir parçasıdır. FBI tarafından biçimlendirilmiş NIBRS verileri, tek bir dosyada saklanmaktadır. Veriler altı farklı kayıt tipi şeklinde organize edilmiştir. Bu kayıt tipleri şunlardır [55]:

- İdari
- Suç
- Mülk
- Kurban
- Suçlu

- Tutuklu

Her kayıt tipi farklı uzunluğa ve düzene sahiptir. Ayrıca, NBRS verileri ile çalışmayı basitleştirmek için dosya versiyonları oluşturulmuştur. Bu versiyonlar aşağıda verilmiştir [55]:

- Olay düzeyinde
- Kurban düzeyinde
- Suçlu düzeyinde
- Tutuklu düzeyinde

NBRS veritabanı kullanıcılara; 1991'den 2014'e kadar toplamda 24 tane, üzerinde analiz yapılabilecek dosya sistemi sunmaktadır [56]. Yapılan tez çalışmasında, 2005 ve 2013 dosyaları kullanılmaktadır. Çalışılan dosya versiyonu ise olay düzeyindedir. Dosyalar içerisinde her bir suç vakası için 378 nitelik bulunmaktadır. Ancak bu niteliklerin hepsi kullanılmamıştır. Yapılan ön işlem süreci sonrasında sadece analizler için gerekli görülen değişkenler kullanılmıştır. Veri tabanında bulunan niteliklerin tamamı EK-1'de verilmiştir.

4.2.2. Küresel Terörizm Veri Tabanı

Küresel Terörizm Veritabanı (GTD), terörist grupların gerçekleştirdiği eylemlerin ayrıntılı bilgisinden oluşan bir veritabanıdır. GTD'de bulunan değişkenlerin tuttukları bilgileri ve bu bilgilerin kodlamalarının ayrıntılı açıklamalarını içeren kodlama dökümanı bulunmaktadır. GTD dökümanı (codebook); Terörizm ve Terörle Mücadeleye Yönelik Ulusal Konsorsiyum (START) tarafından yürütülen, GTD için bilgi toplama ve kodlama kurallarını içermektedir. Bu doküman en son Haziran 2016'da güncellenmiştir [57,58].

GTD toplamda 52135 vaka içermektedir. Her vakanın ayrıntılı bilgisini oluşturan değişkenler ise toplamda 127 dir. Veri tabanında bulunan değişkenlerin tamamı EK-2'de verilmiştir. Bu değişkenler kategorilere ayrıldığında, dokuz kategorinin oluştuğu görülmektedir. Bu kategoriler [58]:

A. GTD id ve saldırı tarihi bilgileri
    a. Her bir saldırı için ayrı bir olay id
    b. Saldırının olduğu gün, ay ve yıl bilgileri
B. Olay bilgileri
    a. Olayın kısa özeti
    b. Saldırının terör saldırısı olma kriterleri
    c. Saldırı terör saldırısı mı? Değilse saldırının tipi
    d. Eylemin birden fazla olayın parçası olup olmadığı
    e. Saldırı ile bağlantılı olaylar.
C. Olayın olduğu lokasyon bilgileri
    a. Toplamda 222 ülke
    b. Ülkeler bulundukları bölgelere göre 12 region'a ayrılmıştır.
        i. Kuzey Amerika Ülkerleri
        ii. Orta Amerika Ülkeleri
        iii. Güney Amerika Ülkeleri
        iv. Doğu Asya Ülkeleri
        v. Güneydoğu Asya Ülkeleri
        vi. Güney Asya Ülkeleri
        vii. Orta Asya Ülkeleri
        viii. Batı Avrupa Ülkeleri
        ix. Doğu Avrupa Ülkeleri
        x. Orta Doğu ve Kuzey Afrika Ülkeleri
        xi. Sahra-altı Afrika Ülkeleri
        xii. Avustralya ve Okyanusya Ülkeleri
    c. Olayın olduğu şehir, ilçe ve tam koordinat bilgileri
D. Saldırı bilgileri

a. Saldırının tipi. Hiyerarşik olarak dokuz atak tipi tanımlanmıştır.
    i. Suikast
    ii. Uçak/Araç Kaçırma
    iii. Adam/Çocuk Kaçırma
    iv. Rehin Alma
    v. Bombalama/Patlama
    vi. Silahlı baskın
    vii. Silahsız saldırı
    viii. Tesis/Altyapı saldırısı
    ix. Bilinmiyor
b. Saldırının başarılı olup olmadığı
c. Saldırının intihar saldırısı olup olmadığı

E. Saldırıda kullanılan silah bilgileri
    a. Saldırıda kullanılan silah tiplerinin en genel kategorize edilmiş hali
        i. Biyolojik silahlar
        ii. Kimyasal silahlar
        iii. Radyolojik silahlar
        iv. Nükleer silahlar
        v. Ateşli silahlar
        vi. Patlayıcılar/bombalar/dinamitler
        vii. Sahte silahlar
        viii. Yangın çıkarıcı silahlar
        ix. Yakın dövüş silahı
        x. Sabotaj ekipmanları
        xi. Diğer
        xii. Bilinmiyor

b. Silahın genel adından özel adını belirten kayıt.

F. Hedef/Kurban bilgileri

a. Hedef/ kurban tipi

i. İş dünyası

ii. Hükümet(genel)

iii. Polis

iv. Askeri

v. Kürtaj kliniklerinde yapılan saldırılar

vi. Hava yollarında yapılan saldırılar

vii. Hükümet(diplomatik)

viii. Eğitim kurumları

ix. Gıda

x. Gazeteciler ve medya

xi. Denizcilik

xii. Sivil toplum örgütleri

xiii. Diğer

xiv. Vatandaş/ gayrimenkul

xv. Din büyükleri/kuruluşlar

xvi. Telekomünikasyon

xvii. Teröristler

xviii. Ulaşım

xix. Bilinmeyen

xx. Kamu hizmetleri

xxi. Şiddet barındıran siyasi partiler

b. Hedef/kurbanın genel adından özel adını belirten kayıt.

c. Saldırılan hedefin ismi

d. Saldırılan hedefin milliyeti

G. Saldırgan bilgileri

a. Saldırgan grubun ismi

b. Saldırganların sayısı

c. Yakalanan saldırganların sayısı

d. Saldırının üstlenip üstlenilmediği

e. Saldırıyı üstlenen grubun bunu duyurma biçimi

f. Birden fazla grubun saldırıyı üstlenip üstlenmediği

g. Olayın altında yatan motivasyon

H. Kayıplar ve sonuçlar

a. Saldırıda olen kişi sayısı

b. Saldırıda ölen saldırgan sayısı

c. Toplam yaralanan kişi sayısı

d. Yaralanan saldırgan sayısı

e. Kaydedilen zaranın maddi boyutu

f. Toplam rehin alınan veya kaçırılan kurban sayısı

g. Kaçırılan kurbanların götürüldüğü yer

h. Kaçırılan veya rehine alınan kurbanların alıkonma süreleri

i. Fidye talebinin olup olmadığı ve parasal değeri

j. Rehine veya kaçırılanların son durumları

i. Kurtarma teşebbüsü

ii. Rehinenin saldırganlar tarafından serbest bırakılması

iii. Rehinenin kaçması

iv. Rehinenin öldürülmesi

v. Başarılı kurtarma

vi. Kombinasyon

vii. Bilinmiyor

İ. Ek bilgiler ve kaynaklar

Veri tabanında toplamda 615 terörist grubun gerçekleştirdiği eylemlerin ayrıntılı bilgileri bulunmaktadır. Ancak birçok terörist grup ile alakalı vaka sayları, grubu tespit edebilmek için yetersizdir. GTD veri tabanı ile yapılacak çalışmada amaç, elde bulunan vakalardan terörist grupları tahmin etmeye çalışmaktır. Bu sebeple ön işlem sürecinde veri tabanından vaka sayısı yetersiz olan terörist grupları çıkarılacak ve analizler kalan gruplar üzerinde gerçekleştirilecektir.

5. SUÇ VAKALARININ ANALİZİ

Suç vakalarının analiz edilmesi, WEKA ve MATLAB platformları üzerinde, iki farklı veri seti kullanılarak gerçekleştirilmiştir. İlk olarak; WEKA platformunda, NIBRS verileri üzerinde suç analizi işlemi gerçekleştirilmiştir. İkinci olarak ise, terörist eylemleri gerçekleştiren grubu tahmin etmeye yönelik bir yazılım geliştirilmiştir. Suç vakalarının analiz edilmesi, genel olarak iki başlık altında incelenmiştir. Bu başlıklar aşağıda verilmiştir:

1. NIBRS ile vakalar üzerinde suç analizi
2. GTD ile terörist grubu tahmin sisteminin oluşturulması

NIBRS ile gerçekleştirilen analizlerde, veri madenciliği aracı olarak WEKA kullanılmıştır. [59]. WEKA platformunun kullanılabilirlik açısından oldukça kullanıcı dostu bir arayüze sahip olması ve literatüre bakıldığında, yapılan birçok çalışmada tercih edilmesi, analizlerde bu aracın kullanılmasına sebep olmuştur. WEKA üzerinde analiz işlemlerinin yapılabilmesi için excel formatında bulunan verilerin, Şekil 5.1' de görüldüğü gibi, ARFF (Attiribute Relation File Format) uzantılı bir dosya formatına çevrilmesi gerekmektedir. Bu formatın yanında WEKA, CSV (Comma Separated Values) formatını da desteklemektedir.

ARFF dosyasını oluştururken ilk olarak ilişki belirtilmeli, sonrasında değişkenler tanımlanmalıdır. Değişken tipleri nominal, nümerik, string vs. olabilmektedir.

```
@relation Tecavuz

@Attribute STATE        NUMERIC
@Attribute B1008        String
@Attribute V20111       NUMERIC
@Attribute V20141       NUMERIC
@Attribute V20171       NUMERIC
@Attribute V30061       NUMERIC
@Attribute V30071       NUMERIC
@Attribute V30081       NUMERIC
```

Şekil 5.1. ARFF dosyasında değişken tanımının yapılması

Değişkenlerin tanımından sonra veri tabanındaki veriler dosyaya eklenmelidir. Bu verileri ARFF dosyasının tanıması için @Data şeklinde bir başlık gerekmektedir. Kayıtlar

arasında ayıraç olarak virgül kullanılmasına dikkat edilmelidir.

GTD kullanılarak gerçekleştirilen yazılımda veri madenciliği aracı olarak MATLAB kullanılmıştır [60]. Oluşturulan yazılımda MATLAB platformunun kullanılmasının sebebi; MATLAB komutlarının, Microsoft Visual Studio'da çalıştırılabilmesidir. MATLAB aracı üzerinde analiz işlemleri excel dosyası üzerinde direk yapılabilmektedir. Yani verileri ayrı bir formata dönüştürmeye gerek kalmamaktadır.

5.1. NIBRS İle Vakalar Üzerinde Suç analizi

NIBRS ile suç vakaları üzerinde üç ayrı analiz işlemi gerçekleştirilmiştir. Gerçekleştirilen bu analizlerden elde edilen bulgular, hedeflenen amaçlar doğrultusunda tartışılmıştır.

5.1.1. Tecavüz vakalarında bilinmeyen unsurların tahmin edilmesi

Yapılan çalışmada; tecavüz vakasında failin olay sırasında veya olaydan kısa bir süre önce uyuşturucu veya alkol kullandığına yönelik bir tahminde bulunulmaya çalışılmıştır. Failin, kurbanın ve olayın genel özellikleri kullanılarak tecavüz verileri üzerinde bir analiz işlemi gerçekleştirilmiştir.

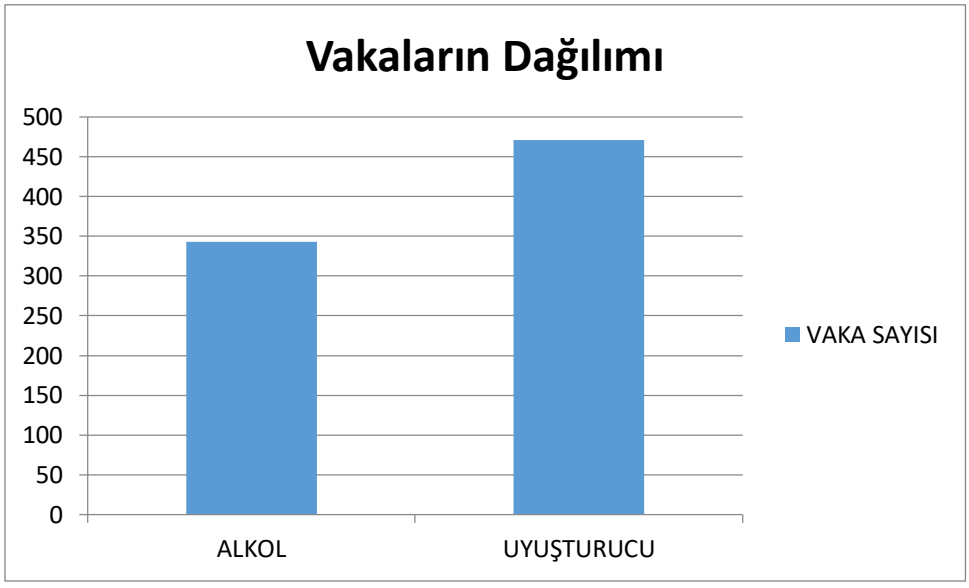
Çalışmada; karar ağaçları, DVM ve NB makine öğrenmesi algoritmaları kullanılarak suç verileri üzerinde analizler gerçekleştirilmiştir. Bu doğrultuda; suç analizinde en uygun algoritmanın belirlenebilmesi için performanslar karşılaştırılmıştır. Suç verileri; FBI tarafından oluşturulan, NIBRS'den elde edilmiştir [61]. Suç tipini oluşturan belirgin unsurlar yardımıyla karanlık taraftaki suç unsurunu yüksek doğrulukla tahmin edilmesini içeren çalışma, bu yönüyle suçların aydınlığa kavuşturulmasında hayati sayılabilecek bilgilerin elde edilmesi açısından önem arz etmektedir.

<u>Suç verilerinin toplanması</u>

Yapılan çalışmada kullanılan tecavüz suçuna ait veriler; FBI tarafından oluşturulan, NIBRS'in 2005 yılına ait verilerinden elde edilmiştir [61].

Veri tabanında bulunan toplam tecavüz vakası 3075, toplam nitelik sayısı ise 378'dir. Bu vakalardan 2575'inde fail alkol, 471'inde fail uyuşturucu kullanmıştır. Kalan 29 vakada ise failin alkol veya uyuşturucu kullandığı şüphesinin bulunmadığı GTD dökümanında belirtilmiştir.

Analizler sonucunda elde edilen sonuçların daha sağlıklı olabilmesi için veri tabanının ön işlem sürecinden geçirilmesi gerekmektedir. Ön işlem sürecinin ilk adımı sınıf dengesizliğini en aza indirgemeye çalışmaktır. Failin alkol kullandığı vaka sayısının uyuşturucu kullandığı vaka sayısından çok fazla olması sınıf dengesizliğine sebep olmaktadır. Bu da yapılan tahminlerin büyük kısmının sayıca fazla olan sınıf tarafında olmasına sebep olmaktadır. Bu doğrultuda, sınıflar arasındaki sayı farkını en aza indirmek için veri tabanından, failin alkol kullandığı vakaların bir kısmı kaldırılmıştır. Ön işlem sürecinin ikinci adımında, veri tabanında eksik verilerin bulunması sebebiyle, toplam nitelik sayısı ve toplam vaka sayısının azaltılmasıdır. Ön işlem sürecinin tamamlanmasından sonra analizler, 343 alkol ve 471 uyuşturucu olmak üzere toplamda 814 tecavüz vakası üzerinde yapılmıştır. Vakaların alkol ve uyuşturucu kullanımına göre dağılımları Şekil 5.2'de verilmiştir.



Şekil 5.2. Vakaların alkol ve uyuşturucu kullanımına göre dağılımları

Önerilen Modelin uygulanması

Çalışmada tecavüz vakasında failin olay sırasında veya olaydan kısa bir süre önce kullandığı maddeyi (uyuşturucu veya alkol) tahmin edebilmek için makine öğrenmesi algoritmalarından NB, karar ağaçları ve DVM kullanılarak WEKA platformu üzerinde bir analiz işlemi gerçekleştirilmiştir.

DVM ile analiz yaparken, ayrım yöntemi olarak polinom (polynomial) kernel kullanılmıştır. Bu kernel, WEKA'da varsayılan olarak ayarlanmıştır. Kullanıcının isteğine göre farklı kernel kullanmak mümkündür.

Karar ağaçları ile sınıflandırma yaparken, güven faktörü (confidence factor) olarak 0.25 değeri kullanılmıştır. Bu değer oluşturulan ağacın budama boyutunu belirlemektedir. Değer 0'a yaklaştıkça budama boyutu küçülürken, 1'e yaklaştıkça budama boyutu büyümektedir. Güven faktörünün WEKA'daki varsayılan değeri 0.25'dir.

Analiz için seçilen niteliklerin suçun aydınlatılmasında yardımcı olabilecek unsurlardan oluşmasına dikkat edilmiştir. Modelin uygulanmasında kullanılan önemli sayılabilecek nitelikler şu şekildedir:

- Eyalet
- Lokasyon
- Silah Tipi
- Kurbanın Cinsiyeti
- Kurbanın Irkı
- Kurbanın Etniği
- Kurbanın İkamet Durumu
- Yaralanma Biçimi
- İlişkili Fail Sayısı
- Fail-Kurban Arasındaki Yakınlık Derecesi
- Failin Yaşı

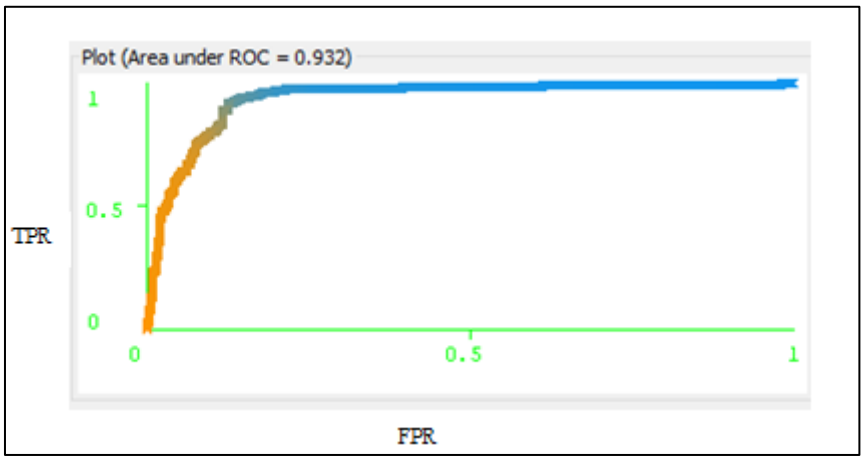
- Failin Cinsiyeti

Yapılan sınıflandırmalarda, öğrenme ve test kümesinin oluşturulmasında 10 katlı çapraz geçerlilik yöntemi kullanılmıştır. Bu yöntemde, öğrenme ve test edilme süreçleri 10 adımda tamamlanmaktadır. Veri kümesinde bulunan tüm veriler öncelikle 10 kümeye ayrılmaktadır. Her adımda kümenin bir kısmı öğrenmeye, kalan kısmı ise teste ayrılmaktadır. Bir sonraki adımda test ve öğrenme kümeleri değiştirilerek, veri kümesindeki tüm verilerin hem test hem de öğrenme için kullanılması sağlanmaktadır. Bu yöntem; öğrenme kümesinin yeterli örnek barındırma olasılığını arttırdığı için, başarım oranını arttırmada katkı sağlayan bir yöntem olarak değerlendirilebilmektedir.

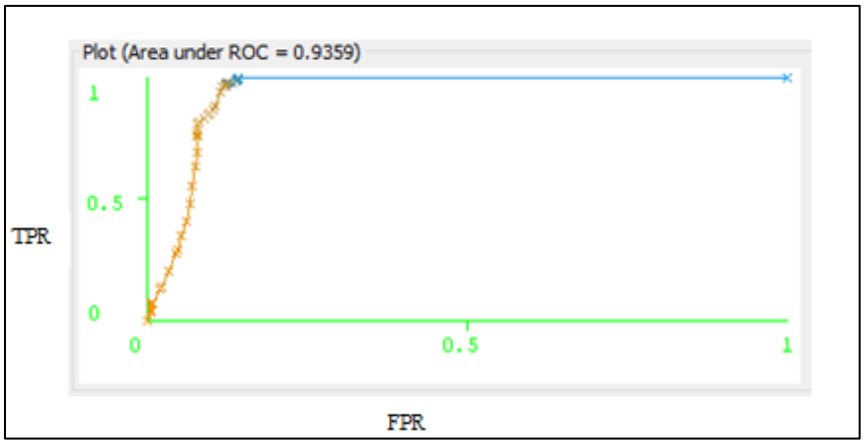
NB, karar ağaçları ve DVM ile oluşturulan modellerin sınıflandırma başarımına ilişkin olarak başarım oranları Çizelge 5.1-5.9' da verilmektedir.

ROC eğrisi, dikey eksen üzerindeki doğru pozitifler (true positive) ve yatay eksen üzerindeki yanlış pozitiflerin (false positive) oranlarının bulunduğu eğridir. Bu eğride, doğru pozitif oranının yüksek, yanlış pozitif oranının düşük olduğu sonuçlar başarılı olarak değerlendirilmektedir. Yani eğri, false positive rate (FPR) eksenine yaklaştıkça başarının seviyesi düşmektedir.

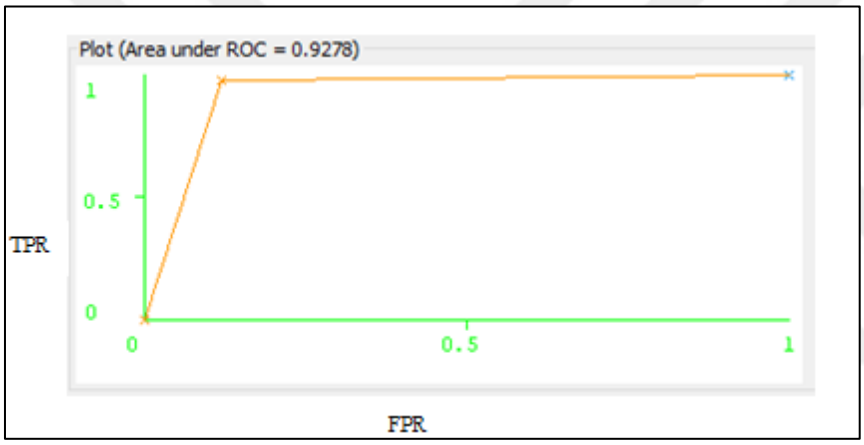
Oluşturulan modellerin ROC (Receiver Operating Characteristic) eğrileri Şekil 5.3-5.5' de gösterilmiştir.



Şekil 5.3. NB sınıflandırıcısı ROC eğrisi



Şekil 5.4. Karar ağaçları sınıflandırıcısı ROC eğrisi



Şekil 5.5. DVM sınıflandırıcısı ROC eğrisi

Çizelge 5.1. NB sınıflandırma modeli başarım oranları

Doğruluk Oranı	Hata Oranı	Kappa İstatistiği	Mutlak Hata	Ortalama Karesel Hata	Göreli Mutlak Hata	Kök Görece Karesel Hata
% 88,083	% 11,916	0,757	0,165	0,305	% 33,974	% 61,933

Çizelge 5.2. NB sınıflandırma modeli karışıklık matrisi

	Tahmin Edilen Sınıf		
Gerçek Sınıf		Sınıf: Alkol	Sınıf: Uyuşturucu
	Sınıf: Alkol	303	40
	Sınıf: Uyuşturucu	57	414

Çizelge 5.3. NB sınıflandırma modeli F-ölçüt değerleri

	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
	0,883	0,121	0,842	0,883	0,862	0,932	Alkol
	0,879	0,117	0,912	0,879	0,895	0,932	Uyuşturucu
Ağırlıklı Ortalama	0,881	0,118	0,882	0,881	0,881	0,932	

Çizelge 5.4. Karar ağaçları sınıflandırma modeli başarım oranları

Doğruluk Oranı	Hata Oranı	Kappa İstatistiği	Mutlak Hata	Ortalama Karesel Hata	Göreli Mutlak Hata	Kök Görece Karesel Hata
% 91,646	% 8,353	0,831	0,128	0,263	%26,289	%53,430

Çizelge 5.5. Karar ağaçları sınıflandırma modeli karışıklık matrisi

		Tahmin Edilen Sınıf	
Gerçek Sınıf		Sınıf: Alkol	Sınıf: Uyuşturucu
	Sınıf: Alkol	333	10
	Sınıf: Uyuşturucu	58	413

Çizelge 5.6. Karar ağaçları sınıflandırma modeli F-ölçüt değerleri

	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
	0,971	0,123	0,852	0,971	0,907	0,936	Alkol
	0,877	0,029	0,976	0,877	0,924	0,936	Uyuşturucu
Ağırlıklı Ortalama	0,916	0,069	0,924	0,916	0,917	0,936	

Çizelge 5.7. DVM sınıflandırma modeli başarım oranları

Doğruluk Oranı	Hata Oranı	Kappa İstatistiği	Mutlak Hata	Ortalama Karesel Hata	Göreli Mutlak Hata	Kök Görece Karesel Hata
% 92,014	% 7.985	0,839	0,079	0,282	%16,374	%52,227

Çizelge 5.8. DVM sınıflandırma modeli karışıklık matrisi

	Tahmin Edilen Sınıf		
Gerçek Sınıf		Sınıf: Alkol	Sınıf: Uyuşturucu
	Sınıf: Alkol	335	8
	Sınıf: Uyuşturucu	57	414

Çizelge 5.9. DVM sınıflandırma modeli F-ölçüt değerleri

	TP Rate	**FP Rate**	**Precision**	**Recall**	**F-Measure**	**ROC Area**	**Class**
	0,977	0,121	0,855	0,977	0,912	0,928	Alkol
	0,879	0,023	0,981	0,879	0,927	0,928	Uyuşturucu
Ağırlıklı Ortalama	0,92	0,064	0,928	0,92	0,921	0,928	

Çizelge 5.1-5.9' da görüldüğü gibi yapılan sınıflandırmalar sonucunda en yüksek doğruluk oranına %92 ile DVM ulaşmıştır. En düşük doğruluk oranı ise %88 ile NB algoritması olmuştur. Karar ağaçlarının başarım oranı ise %91 olmuştur. 814 tecavüz vakası içinde NB 717, karar ağaçları 746 ve DVM 749 vakada failin kullandığı maddeyi doğru tahmin edebilmiştir.

ROC eğrisi, false positive rate (FPR) eksenine yaklaştıkça başarının seviyesi düşmektedir. Bu doğrultuda DVM sınıflandırıcısının ROC eğrisi; doğruluk oranının en yüksek olması ile doğru orantılı olarak, en başarılı grafiktir.

Vakada bilinmeyen unsurların ortaya çıkarılması, işlenen suçların aydınlığa kavuşturulması açısından son derece önem arz etmektedir. Suç olayında, önemsiz olarak nitelendirilen bir olgu, o vakanın çözümünde kilit nokta olabilmektedir. Bu sebeple işlenen suçun her yönünün açık ve net olması, vakanın çözülmesi açısından önemlidir. Çalışmada; tecavüz suçuna ait 814 vaka üzerinde, bilinmeyen unsurların tahmin edilebilmesi için analiz yapılmıştır. Tahminlerin %90'a varan yüksek doğruluğu, suç analizinde makine öğrenmesi kullanımının, başarılı olabilecek bir yaklaşım olduğu sonucunu doğurmaktadır.

5.1.2. Cinayet vakalarında gizli ilişkilerin keşfedilmesi

Çalışmanın bu kısmında; veri madenciliği modellerinden birliktelik kuralları ve bu kuralların üretilmesi için apriopri algoritması kullanılarak, suçu oluşturan unsurlar arasındaki ilişkiler ve birliktelikler bulunmaya çalışılmıştır.

Apriopri algoritması, bağlantı analizlerinin yapılıp bağlantı kurallarının ortaya çıkartılması konusunda en çok bilinen ve kullanılan algoritmadır [31]. Bu algoritma kullanılarak, 3142 cinayet vakası üzerinde bir analiz işlemi gerçekleştirilmiştir. Birliktelik kuralları modelinin ve apriopri algoritmasının uygulanmasında suç, suçlu ve kurbana ait çeşitli niteliklerden faydalanılmıştır. Bu nitelikler; suçun işlendiği saat, suçun işlenmesinde kullanılan silah tipi, suçla ilişkili olabilecek fail sayısı, kurbanın yaralanma biçimi, kurbanın ve failin yaşı, cinsiyeti, ırkı, etnik kökeni gibi, suçun analiz edilmesinde önemli sayılabilecek niteliklerdir.

Suç verilerinin toplanması

Çalışmada kullanılan veri, NIBRS veri tabanındaki 2013 kayıtlarından elde edilmiştir [62]. Çalışılan dosya versiyonu ise olay düzeyinde dosyadır. Dosya içerisinde toplamda 46 suç tipine ait veriler bulunmakta ve her bir suç vakası için 378 nitelik bulunmaktadır. Yapılan çalışmada 3142 cinayet vakası üzerinde çalışılmış ve nitelik olarak ise 378 nitelik içinden 16 nitelik birliktelik kuralları modelini uygulamak için kullanılmıştır.

Önerilen Modelin uygulanması

Cinayet vakalarında bilinmeyen ilişkilerin ortaya çıkarılması için; NBRS veri tabanındaki 3142 vaka üzerinde çalışılarak bir analiz işlemi gerçekleştirilmiştir. Analiz işleminde failin, mağdurun ve olayın genel özelliklerini içeren 16 nitelik kullanılmıştır. Analiz sonucunda beklenen; niteliklerin birbiriyle olan bağlantılarını açıklayan veya niteliklerin birlikteliklerinin olasılıklarını ortaya koyan birliktelik kurallarının oluşturulması ve bu doğrultuda cinayet vakalarındaki gizli ilişkilerin keşfedilmesidir. Çalışmada önerilen modelin uygulanması için ücretsiz ve açık kaynak kodlu bir yazılım olan WEKA aracı kullanılmıştır. Analizde kullanılan nitelikler Tablo 1'de verilmiştir. Modelin uygulanması sonucunda elde edilen ilk 10 birliktelik kuralı ise Şekil 5.6' de verilmiştir.

```
Best rules found:

 1. KurbaninEtnikKokeni=0 KurbaninIkametDurumu=1 FailinCinsiyeti=1 1283 ==> IliskiliOlabilecekFailSayisi=1 1248    conf:(0.97)
 2. KurbaninEtnikKokeni=0 FailinCinsiyeti=1 1579 ==> IliskiliOlabilecekFailSayisi=1 1532    conf:(0.97)
 3. KurbaninIkametDurumu=1 FailinCinsiyeti=1 1685 ==> IliskiliOlabilecekFailSayisi=1 1631    conf:(0.97)
 4. KurbanCinsiyeti=1 KurbaninIkametDurumu=1 FailinCinsiyeti=1 1183 ==> IliskiliOlabilecekFailSayisi=1 1145    conf:(0.97)
 5. Lokasyontipi=20 FailinCinsiyeti=1 1229 ==> IliskiliOlabilecekFailSayisi=1 1181    conf:(0.96)
 6. FailinCinsiyeti=1 2229 ==> IliskiliOlabilecekFailSayisi=1 2135    conf:(0.96)
 7. KurbanCinsiyeti=1 FailinCinsiyeti=1 1601 ==> IliskiliOlabilecekFailSayisi=1 1533    conf:(0.96)
 8. FailinIrki=2 1292 ==> IliskiliOlabilecekFailSayisi=1 1236    conf:(0.96)
 9. FailinCinsiyeti=1 FailinIrki=2 1199 ==> IliskiliOlabilecekFailSayisi=1 1145    conf:(0.95)
10. SuçluAktiviteninTipi=7 FailinCinsiyeti=1 1387 ==> IliskiliOlabilecekFailSayisi=1 1320    conf:(0.95)
```

Şekil 5.6. Modelin uygulanması ile elde edilen birliktelik kuralları

Cinayet verileri üzerine birliktelik kuralları modelinin ve apriopri algoritmasının uygulanması sonucu elde edilen birliktelik kuralları aşağıdaki gibi yorumlanmıştır:

1. Kurban ispanyol veya latin kökenli ve cinayet kurbanın sürekli oturduğu evde gerçekleşti ise, büyük olasılıkla cinayetle bağlantılı fail sayısı 1'dir.
2. Kurban ispanyol veya latin kökenli ve birincil failin cinsiyeti erkek ise, büyük olasılıkla cinayetle bağlantılı fail sayısı 1'dir.
3. Cinayet kurbanın sürekli oturduğu evde gerçekleşti ve birincil failin cinsiyeti erkek ise, büyük olasılıkla cinayetle bağlantılı fail sayısı 1'dir.
4. Kurbanın cinsiyeti erkek ve cinayet kurbanın sürekli oturduğu evde gerçekleşti ise, büyük olasılıkla cinayetle bağlantılı fail sayısı 1'dir.
5. Cinayet evde işlenmiş ve birincil failin cinsiyeti erkek ise, büyük olasılıkla cinayetle bağlantılı fail sayısı birdir.
6. Birincil failin cinsiyeti erkek ise, büyük olasılıkla cinayetle bağlantılı fail sayısı birdir.
7. Kurbanın cinsiyeti ve birincil failin cinsiyeti erkek ise, büyük olasılıkla cinayetle bağlantılı fail sayısı birdir.
8. Birincil fail siyahi veya afro-amerikan ise, büyük olasılıkla cinayetle bağlantılı fail sayısı birdir.
9. Birincil failin cinsiyeti erkek ve birincil fail siyahi veya afro-amerikan ise, büyük olasılıkla cinayetle bağlantılı fail sayısı birdir.
10. Cinayet bilinmeyen çete tarafından işlendi ve birincil fail erkek ise, büyük olasılıkla cinayetle bağlantılı fail sayısı 1'dir.

Yapılan çalışmada veri madenciliği modellerinden birliktelik kuralları ve bu kuralların üretilmesi için apriopri algoritması kullanılarak, 3142 cinayet vakası üzerinde bir analiz

işlemi gerçekleştirilmiştir. Uygulanan modelin ve yapılan analizin sonucunda elde edilen ilk 10 birliktelik kuralı yorumlanmıştır. Oluşturulan birliktelik kurallarının yorumlanması; suç ve suçu oluşturan unsurlar arasındaki ilişkileri anlamada yardımcı olacak ve bu doğrultuda suçun çözümüne ışık tutabilmesi açısından güvenlik güçlerine fayda sağlayacaktır.

5.1.3. Suç vakasının tahmin edilmesi

Çalışmada; suç türünün tahmin edilebilmesi amacıyla; failin, mağdurun veya olayın genel özelliklerinden faydalanılarak cinayet, tecavüz ve adam kaçırma verileri üzerinde bir analiz işlemi gerçekleştirilmiştir. Makine öğrenmesi algoritmalarından biri olan YSA, modelin uygulanmasında kullanılmıştır. Sınıf dengesizliğini çözümleyebilmek için ise, önyargılı (biased) örneklem dağılımı yaklaşımı uygulanmıştır. Analizde kullanılan suç verileri; FBI tarafından oluşturulan, NIBRS'den elde edilmiştir. Yapılan analizler; açık kaynak kodlu, ücretsiz bir yazılım olan WEKA platformu üzerinde gerçekleştirilmiştir.

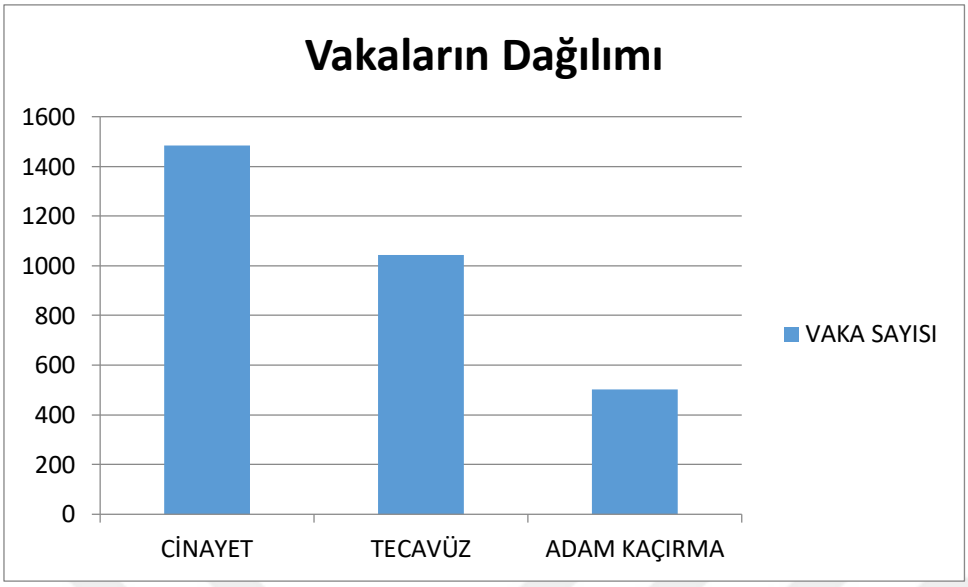
Suç verilerinin toplanması

Analizde kullanılan veriler, NIBRS'in 2013 kayıtlarından elde edilmiştir. Suç olaylarının geneli üzerinde bir analiz işlemi gerçekleştirildiğinden, dosya versiyonu olarak olay düzeyinde dosya tercih edilmiştir. Çalışmada veri tabanında bulunan 46 suç tipi içerisinden; Tecavüz, Adam kaçırma ve Cinayet suç tipleri, yapılan sınıflandırmada, sınıf etiketi olarak kullanılmıştır. 1500 cinayet, 1036 tecavüz ve 494 adam kaçırma suçlarına ilişkin veriler üzerinde çalışılmıştır. Nitelik olarak ise 378 nitelik içinden 22 nitelik sınıflandırma için kullanılmıştır. Suç tiplerini ayırt etmede önemli sayılabilecek niteliklerin analizde kullanılmasına dikkat edilmiştir.

Analizde kullanılan vakaların; tecavüz, adam kaçırma ve cinayet suç tiplerine göre dağılımları Şekil 5.7' de verilmektedir. Sınıflandırmada kullanılan nitelikler ise şu şekildedir:

- Suçun işlendiği Eyalet
- Suçun İşlendiği Saat

- Failin Suçu İşlemeden Önce Kullandığı Madde
- Suçun İşlendiği Lokasyon Tipi
- Suçlu Aktivitenin Tipi
- Suçta Kullanılan Silah Tipi
- Failin Suçu İşlemedeki Motivasyonu
- Suç Sırasında Kaybolan Eşya
- Kayıp Eşyanın Tanımı
- Kayıp Eşyanın Değeri
- Kurban Tipi
- Kurbanın Cinsiyeti
- Kurbanın Irkı
- Kurbanın Etniği
- Kurbanın Yerleşim Durumu
- Kurbanın Yaralanma Biçimi
- Suçla ilişkili Olabilecek Fail Sayısı
- Kurban ve Fail Arasındaki Akrabalık İlişkisi
- Failin Yaşı
- Failin Cinsiyeti
- Failin Irkı



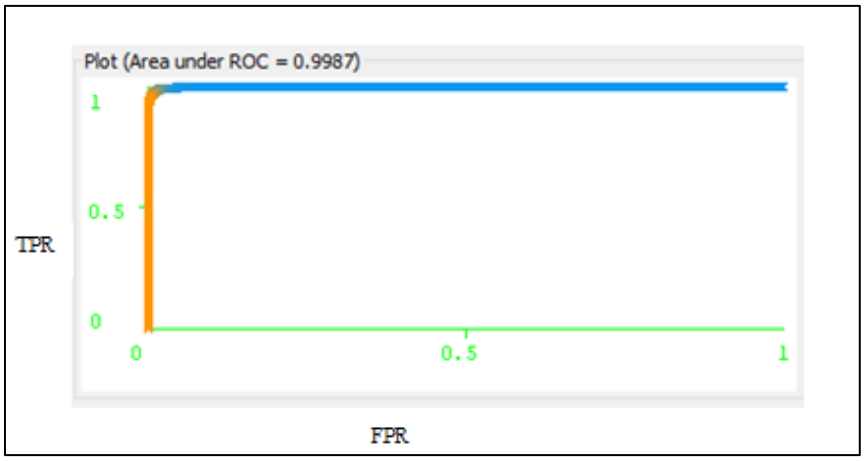
Şekil 5.7. Vakaların cinayet, tecavüz ve adam kaçırma suç tiplerine göre dağılımları

Önerilen Modelin uygulanması

Önerilen modelin uygulanmasında ilk olarak; seçilen algoritma veri kümesi üzerine uygulanmış, daha sonra önyargılı örneklem dağılımı yaklaşımı ile analiz tekrarlanmıştır. Son olarak, bu iki analiz sonucunda elde edilen performans çıktıları karşılaştırılmıştır.

Cinayet, tecavüz ve adam kaçırma verilerinin sınıflandırılması

Yapılan sınıflandırmada, öğrenme ve test kümesinin oluşturulmasında 10 katlı çapraz geçerlilik yöntemi kullanılmıştır. YSA ile oluşturulan modelin sınıflandırma başarımına ilişkin olarak başarım oranları Çizelge 5.10-5.12' da verilmektedir. Ayrıca Oluşturulan modelin ROC (Receiver Operating Characteristic) eğrisi Şekil 5.8' de gösterilmiştir.



Şekil 5.8. YSA sınıflandırıcısı ROC eğrisi

Çizelge 5.10. YSA sınıflandırma modeli başarım oranları

Doğruluk Oranı	Hata Oranı	Kappa İstatistiği	Mutlak Hata	Ortalama Karesel Hata	Göreli Mutlak Hata	Kök Görece Karesel Hata
% 97.029	% 2.970	0,951	0,022	0,128	% 5,521	% 28,420

Çizelge 5.11. YSA sınıflandırma modeli karışıklık matrisi

	Tahmin Edilen Sınıf			
Gerçek Sınıf		Sınıf: Cinayet	Sınıf: Tecavüz	Sınıf: Adam k.
	Sınıf: Cinayet	1463	16	7
	Sınıf: Tecavüz	17	1005	21
	Sınıf: Adam kaçırma	10	19	472

Çizelge 5.12. YSA sınıflandırma modeli F-ölçüt değerleri

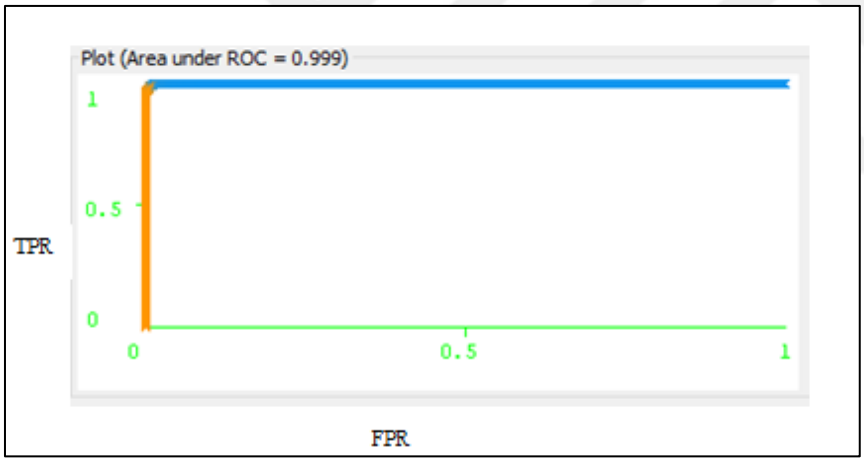
	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
	0,985	0,017	0,982	0,985	0,983	0,999	Cinayet
	0,964	0,018	0,966	0,964	0,965	0,995	Tecavüz
	0,942	0,011	0,944	0,942	0,943	0,992	Adam K.
Ağırlıklı Ortalama	0,97	0,016	0,97	0,97	0,97	0,966	

Doğruluk oranına bakıldığında %97 gibi bir değerle karşılaşılmaktadır. Bu oran algoritma başarısı açısından değerlendirildiğinde oldukça iyidir. Ancak karışıklık matrisi

yorumlandığında, cinayet ve tecavüz vaka sayılarının çoğunluk olması sebebiyle, doğru suç tipi farklı olmasına rağmen, tahminler en fazla bu sınıflar doğrultusunda yapılmıştır. Bu durum hata oranının yüksek olmasına sebep olmuştur.

Cinayet, tecavüz ve adam kaçırma verilerinin önyargılı örneklem dağılımı yaklaşımı ile sınıflandırılması

Önyargılı örneklem dağılımı yaklaşımının YSA'da kullanılması ile oluşturulan modelin sınıflandırma başarımına ilişkin olarak başarım oranları Şekil 5.13-5.15' de verilmektedir. Ayrıca Oluşturulan modelin ROC eğrisi Şekil 5.9' da gösterilmiştir. Önyargılı örneklem dağılımı yaklaşımının YSA'ya uygulanmasıyla, doğruluk oranının %98 olduğu görülmektedir. İlk analiz sonuçlarıyla karşılaştırıldığında, bu yaklaşımın başarım oranını arttırdığı görülmektedir.



Şekil 5.9. Önyargılı örneklem dağılımı yaklaşımı ile YSA sınıflandırıcısı ROC eğrisi

Çizelge 5.13. Önyargılı örneklem dağılımı ile YSA modeli başarım oranları

Doğruluk Oranı	Hata Oranı	Kappa İstatistiği	Mutlak Hata	Ortalama Karesel Hata	Göreli Mutlak Hata	Kök Görece Karesel Hata
% 98,41	% 1,28	0,974	0,012	0,093	% 3,173	% 20,799

Çizelge 5.14. Önyargılı örneklem dağılımı ile YSA modeli karışıklık matrisi

		Tahmin Edilen Sınıf		
Gerçek Sınıf		Sınıf: Cinayet	Sınıf: Tecavüz	Sınıf: Adam k.
	Sınıf: Cinayet	1482	15	3
	Sınıf: Tecavüz	12	1016	8
	Sınıf: Adam kaçırma	3	7	484

Çizelge 5.15. Önyargılı örneklem dağılımı ile YSA modeli F-ölçüt değerleri

	TP Rate	**FP Rate**	**Precision**	**Recall**	**F-Measure**	**ROC Area**	**Class**
	0,988	0,01	0,99	0,988	0,989	0,999	Cinayet
	0,981	0,011	0,979	0,981	0,98	0,996	Tecavüz
	0,98	0,004	0,978	0,98	0,979	0,998	Adam K.
Ağırlıklı Ortalama	0,984	0,009	0,984	0,984	0,984	0,998	

Suç vakasını oluşturan unsurların bilinmesi; suç, suçlu ve kurban arasındaki bağlantıları anlamada ve bu doğrultuda suçun aydınlığa kavuşturulmasında önemlidir. NIBRS bu anlamda; suç vakaları üzerinde çeşitli analizlere imkan sağlaması açısından kullanılabilecek veritabanlarından bir tanesidir. Çalışmada; tecavüz, cinayet ve adam kaçırma vakaları üzerinde, YSA ile suç türünün tahmin edilmesine çalışılmıştır. Ek olarak sınıf dengesizliğini gidermek için önyargılı örneklem dağılımı yaklaşımı kullanılmış ve daha doğru sınıflandırma yapılması amaçlanmıştır. Bu doğrultuda; önyargılı örneklem dağılımının sınıf dengesizliğinin çözümünde kullanılması, YSA'nın başarım oranını arttırması açısından faydalı bir yaklaşım olmuştur.

5.2. GTD İle Terörist Grubu Tahmin Sistemi

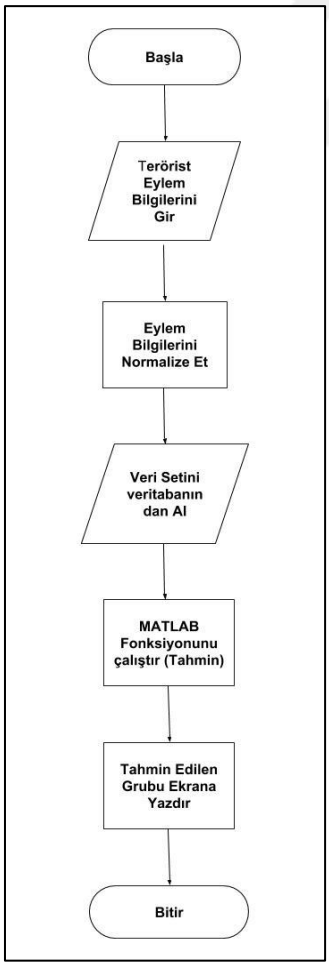
Geliştirilen uygulamada 3 platform kullanılmıştır.

- Microsoft SQL Server Management Studio
- Matlab
- Microsoft Visual Studio

Eğitim ve test kümeleri Microsoft SQL Server Management Studio'da tutulmaktadır. Microsoft Visual Studio kullanıcı arayüzü oluşturulması kısmını oluştururken, ağın eğitilmesi için Matlab kullanılmaktadır. Terörist eylemlerin vakaları üzerinden gerçekleştirilen terörist grubu tahmin sisteminin oluşturulma aşamaları sırasıyla şu şekildedir:

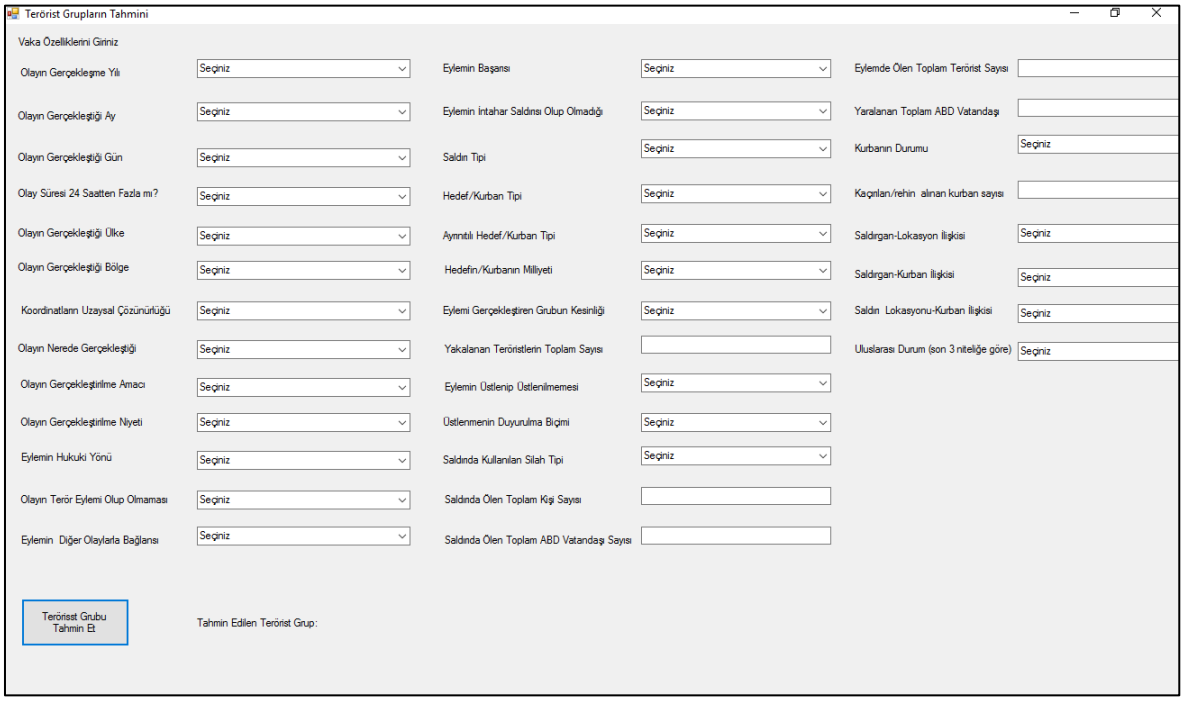
1. Veritabanı üzerinde yapılan ön işlem süreci
2. Eğitim ve test kümelerinin oluşturulması
3. Eğitim ve test kümelerinin veri tabanına aktarılması
4. Matlab'da gerçekleştirilen işlemler
5. Microsoft Visual Studio' da gerçekleştirilen işlemler

Terörist grubu tahmin sisteminin akış diyagramı Şekil 5.10' de verilmektedir.



Şekil 5.10. Terörist grubu tahmin sistemi akış diyagramı

Akış diyagramından da görüldüğü gibi; sistem ilk olarak kullanıcıdan terörsit grubu tahmin edilecek vaka bilgilerini istemektedir. Daha sonra bu bilgiler; eğitilmiş ve öğrenmiş ağda kullanılabilmesi için normalize edilmektedir. Veri tabanından MATLAB fonksiyonunun çalıştırılabilmesi için eğitim ve test kümeleri veri tabanından çekilmekte ve bu kümeler, MATLAB fonksiyonuna parametre olarak gönderilmektedir. Son işlem olarak; bu fonksiyonun çalıştırılmasıyla, eldeki vakanın hangi terörist grubun eylemi olduğu tahmin edilmektedir. Terörist grubu tahmin sistemin kullanıcı arayüzü Şekil 5.11' da verilmiştir.



Şekil 5.11. Kullanıcı arayüzü

5.2.1. Veritabanı üzerinde yapılan ön işlem süreci

Veritabanında her vakanın ayrıntılı bilgisini oluşturan değişkenler toplamda 127 dir. Yapılacak olan analizlerin daha sağlıklı olabilmesi için, ön işlem sürecinde bu değişkenlerin bir kısmı veri tabanından çıkarılmıştır. Son durumda veritabanında 34 değişken bulunmaktadır. Bu değişkenler aşağıda verilmektedir:

- Olayın gerçekleşme yılı
- Olayın gerçekleştiği ay
- Olayın gerçekleştiği gün

- Olay süresi 24 saatten fazla mı?
- Olayın gerçekleştiği ülke
- Olayın gerçekleştiği bölge
- Koordinatların uzaysal çözünürlüğü
- Olayın nerede gerçekleştiği
- Olayın gerçekleşme amacı
- Olayın gerçekleşme niyeti
- Eylemin hukuki yönü
- Olayın terör eylemi olup olmaması
- Eylemin diğer olaylarla bağlantısı
- Eylemin başarısı
- Eylemin intihar saldırısı olup olmadığı
- Saldırı tipi
- Hedef/kurban tipi
- Ayrıntılı hedef/kurban tipi
- Hedef/kurban milliyeti
- Terörist grubun adı
- Eylemi gerçekleştiren grubun kesinliği
- Yakalanan teröristlerin toplam sayısı
- Eylemin üstlenip üstlenilmemesi
- Üstlenmenin duyurulma biçimi
- Saldırıda kullanılan silah tipi
- Saldırıda ölen toplam kişi sayısı
- Saldırıda ölen toplam ABD vatandaşı
- Eylemde ölen toplam terörist sayısı

- Yaralanan toplam ABD vatandaşı
- Kurbanların Durumu
- Kaçırılan/rehin alınan kurban sayısı
- Saldırgan/lokasyon ilişkisi
- Saldırgan/kurban ilişkisi
- Saldırı lokasyonu/kurban ilişkisi
- Ulusal veya uluslararası durum

Ön işlem sürecinde veritabanında kullanılan değişkenler, analizde kullanılabilecek hale getirilmeye çalışılmıştır. Değişkenler üzerinde yapılan ön işlem sürecinde karşılaşılan önemli sayılabilecek durumlar aşağıda verilmiştir.

Gereksiz değişkenlerin veritabanından çıkartılması:

Veritabanında zaten bulunan bilgilerin kaydını tutan değişkenler veritabanından çıkarılmıştır.

Fazla ayrıntı içeren değişkenlerin veritabanından çıkartılması:

Bazı değişkenler, vaka ile ilgili fazla ayrıntılı bilgi içerdiğinden veritabanından çıkartılmıştır. Bu değişkenlerin daha genel halleri zaten veritabanında bulunmaktadır. Fazla ayrıntılı bilgi, yanlış öğrenmeye sebep olabileceğinden bu değişkenler veri tabanından çıkarılmıştır.

Boş hücre sayısının dolu hücre sayısından fazla olan değişkenlerin veri tabanından çıkartılması:

Bazı değişkenlerde bulunan hücrelerin çok az bir kısmı kayıt tuttuğu için bu değişkenler veri tabanından çıkartılmıştır.

Veritabanında bulunan birbirine bağımlı değişkenler:

targtype1=20 ise hedef tipi belli değildir. Bu sebepten targsubtype1 de belli değildir. Belli olmayan targsubtype1'lere veritabanında bu değişken tipi için kullanılmayan bir kodlama yapılması gerekmektedir. targtype1=20 ise targsubtype1= 112 şeklinde bir kodlama yapılmıştır.

targsubtype1'de boş olan hücreler için veritabanı raporunun yorumu: "if a target subtype is not applicable this variable is left blank." Yani, bazı hedeflerin alt tipi bulunmamaktadır. Bu sebeple targsubtype1 boş olan hücrelerde alt tip uygulanmamıştır. Bu durumda yeni bir kodlamaya ihtiyaç duyulmuştur. targsubtype1'de boş olan hücreler= 113 ile kodlanmıştır.

Claimed=0 olduğu durumlarda claimmode boş bırakılmıştır. Çünkü claimmode, claimed'in 0 olduğu durumlarda uygulanmamaktadır. Claimed=0 ise claimmode=11 şeklinde kodlama yapılmıştır.

İshostkid değişkeninin 0 olduğu durumlarda nhostkid değeri boş bırakılmıştır. İshostkid=0 ise nhostkid=0 kodlaması yapılmıştır.

Ön işlem sürecinde ikinci adım kullanılmayacak veriyi veritabanından kaldırmaktır. Veri tabanında bulunan 615 terörist grubun arasından analiz için vaka sayısı yetersiz olan gruplar, veri tabanından çıkarılmıştır. Çalışmada kullanılacak terörist grupların isim, sayıları ve dağılımları Çizelge 5.16. ve Şekil 5.12'de verilmiştir.

Çizelge 5.16. Terörist grupların toplam eylem sayıları

	Toplam Eylem Sayısı
Al-Shabaab	1295
Al-Qaida in the Arabian Peninsula (AQAP)	716
BokoHaram	1449
New People's Army (NPA)	858
PKK	521
Taliban	1500
Tehrik-i-Taliban Pakistan (TTP)	584

Şekil 5.12. Terörist grupların dağılımı

Şekil 5.12' de görüldüğü gibi analiz işlemlerinde kullanılmak üzere yedi terörist grup belirlenmiştir.

Veriler üzerinde ön işlem yapılmasındaki son işlem, veri dönüştürmedir. Oluşturulacak olan eğitim ve test kümelerinin matlab'da kullanılabilmesi için normalizasyon işlemi gerçekleştirmek gerekmektedir. Bu amaçla min-max normalizasyonu, tüm veriler üzerine uygulanmıştır. Böylelikle tüm veriler 0-1 değerleri arasında değerlere sahip olmuştur.

5.2.2. Eğitim ve test kümelerinin oluşturulması

Ön işlem sürecinin tamamlanmasından sonraki adım eğitim ve test kümelerinin oluşturulmasıdır. Eğitim ve test kümelerinin oranları, analizlerin başarısı açısından oldukça önemlidir. Eğitim kümesinin problem için yeterince örneklem içermemesi eğitim tam olarak yapılamamasına, bunun sonucu başarım oranının düşüklüğüne sebep olmaktadır. Eğitim kümesinin gereğinden fazla örneklemle eğitilmesi durumunda ise ezberleme dediğimiz durum ortaya çıkmakta ve başarım oranı yine düşmektedir. Literatüre bakıldığında eğitim ve test kümelerinin oranları ile ilgili çok fazla çalışma vardır. Genel kanı eğitim kümesinin %66 ve test kümesinin %34 alındığında optimum sonuçlara ulaşılabileceğidir. Bu doğrultuda terörist grupların vakalarından oluşturulan eğitim ve test kümeleri Çizelge 5.17'de verilmektedir.

Çizelge 5.17. Terörist grupların eğitim ve test kümesi olarak dağılımı

	Eğitim Kümesi	Test Kümesi
Al-Shabaab	855	440
Al-Qaida in the Arabian Peninsula (AQAP)	473	243
BokoHaram	956	493
New People's Army (NPA)	566	292
PKK	344	177
Taliban	990	510
Tehrik-i-Taliban Pakistan (TTP)	385	199

5.2.3. Eğitim ve test kümelerinin veritabanına aktarılması

Eğitim ve test kümelerinin uygulamada kullanılabilmesi için bu kümelerin, öncelikle SQL Server Management Studio'ya aktarılması gerekmektedir. Buraya aktarılan eğitim ve test kümeleri daha sonra Microsoft Visual Studio'dan çağırılacak ve Matlab kullanılarak ağ eğitme işlemi gerçekleştirilecektir.

İlk olarak SQL Server Management Studio'da "TerörizmVeriseti" adında bir veri tabanı oluşturulup; eğitim ve test kümelerini barındıracak olan tablolar TerörizmVeriseti veritabanına eklenmiştir. Bu tablolara excel formatında olan veriler import edilmiş, böylelikle uygulamada kullanılacak eğitim ve test kümeleri oluşturulmuştur.

5.2.4. Yapay sinir ağında eğitim ve testlerin gerçekleştirilmesi

Yapay sinir ağının eğitilme kısmı MATLAB'da oluşturulan bir fonksiyon aracılığıyla gerçekleştirilmiştir. Bu fonksiyon daha sonra uygulama ara yüzünde çağırılacak ve ardından ağ eğitilecektir. MATLAB'da gerçekleştirilen işler sırasıyla şunlardır:

1. Yapay sinir ağının oluşturulması ve ağın eğitilmesi
2. Eğitilen ağ üzerinde test işleminin gerçekleştirilmesi

Yapay sinir ağının oluşturulması ve ağın eğitilmesi

Yapay sinir ağının oluşturulabilmesi için öncelikle; ağ bilgilerini saklayacak yapı olan ağ objesinin oluşturulması gerekmektedir. Bu ağ objesinin girdileri şunlar olmaktadır:

- Girdi
- Çıktı
- Gizli katmanlardaki nöron sayıları
- Aktivasyon fonksiyonu olarak ne kullanılacağı
- Eğitim algoritması olarak ne kullanılacağı

Daha sonra yine bu ağ için öğrenme parametreleri belirlenmelidir. Bu parametreler; ağın öğrenme katsayısı, öğrenmenin iterasyon sayısının ayarlanması (epoch sayısı) gibi parametrelerdir. Öğrenme parametrelerinin de belirlenmesiyle birlikte, eğitime hazır ileri beslemeli bir ağ oluşturulmuştur.

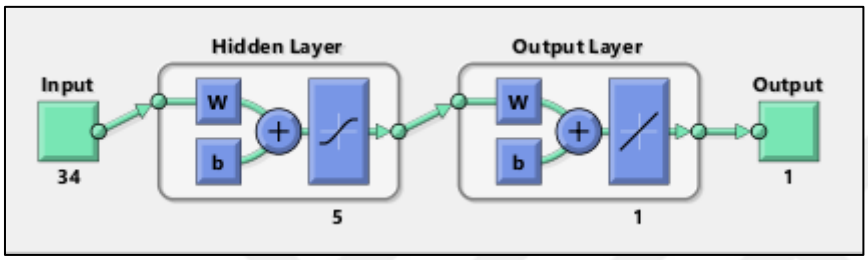
Yapay sinir ağının eğitilmesinde; danışmanlı öğrenme yöntemi kullanılmaktadır. Yani ağa verilen giriş değerlerinin çıktı değerleri de bellidir.

Matlab'da oluşturulan yapay sinir ağında; gizli katman sayısı, bu katmanlardaki nöron sayları, aktivasyon fonksiyonu, eğitim algoritması ve öğrenme parametreleri değiştirilebilir yapıda tasarlanmıştır. En iyi öğrenmenin gerçekleştirilmesi hedeflendiğinden, bu değerler değiştirilerek analizler tekrarlanmış ve sonuçlar tartışılmıştır.

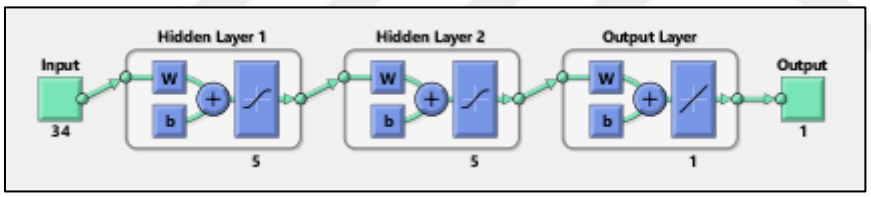
Eğitilen ağ üzerinde test işleminin gerçekleştirilmesi

Eğitimi tamamlanmış ağın başarısını ölçülmesi için, eğitilmiş ağa eğitimde kullanılmayan örnekler verilmeli ve yeni gelen örneklere karşı davranışı incelenmelidir. İyi bir yapay sinir ağı minimum hataya sahiptir. Ağdaki hata, hata ölçüm fonksiyonları ile ölçülmektedir. Hata fonksiyonlarına ek olarak ağın başarısının ölçülmesinde kullanılabilecek diğer bir yaklaşım ise doğruluk oranının hesaplanmasıdır. Bu yaklaşım; toplam verilen test kümesi içinde kaç vakanın doğru, kaç vakanın yanlış tahmin edildiğini göstermektedir. Böylelikle yüzde olarak bir başarıdan söz etmek mümkün olacaktır.

Yapılan analizler iki aşamadan oluşmaktadır. Birinci aşamada tek katmanlı, ikinci aşamada ise iki katmanlı sinir ağları modellenmiştir. Modellenen sinir ağları Şekil 5.13 ve Şekil 5.14'de gösterilmektedir. Her analizde, nöron sayıları değiştirilerek deneyler tekrarlanmıştır. Deneylerin tekrarlanmasının sebebi; ağın hem katman sayısına, hem de nöron sayısına olan yaklaşımını ölçebilmektir. Bu analizler sonucunda ağlardan elde edilen performanslar Çizelge 5.18 ve 5.19'da verilmektedir. Yine ağın performans ölçümü için kullanılan Mean squared error (mse) hata fonksiyonunun grafikleri Şekil 5.15 ve 5.16'de verilmektedir.



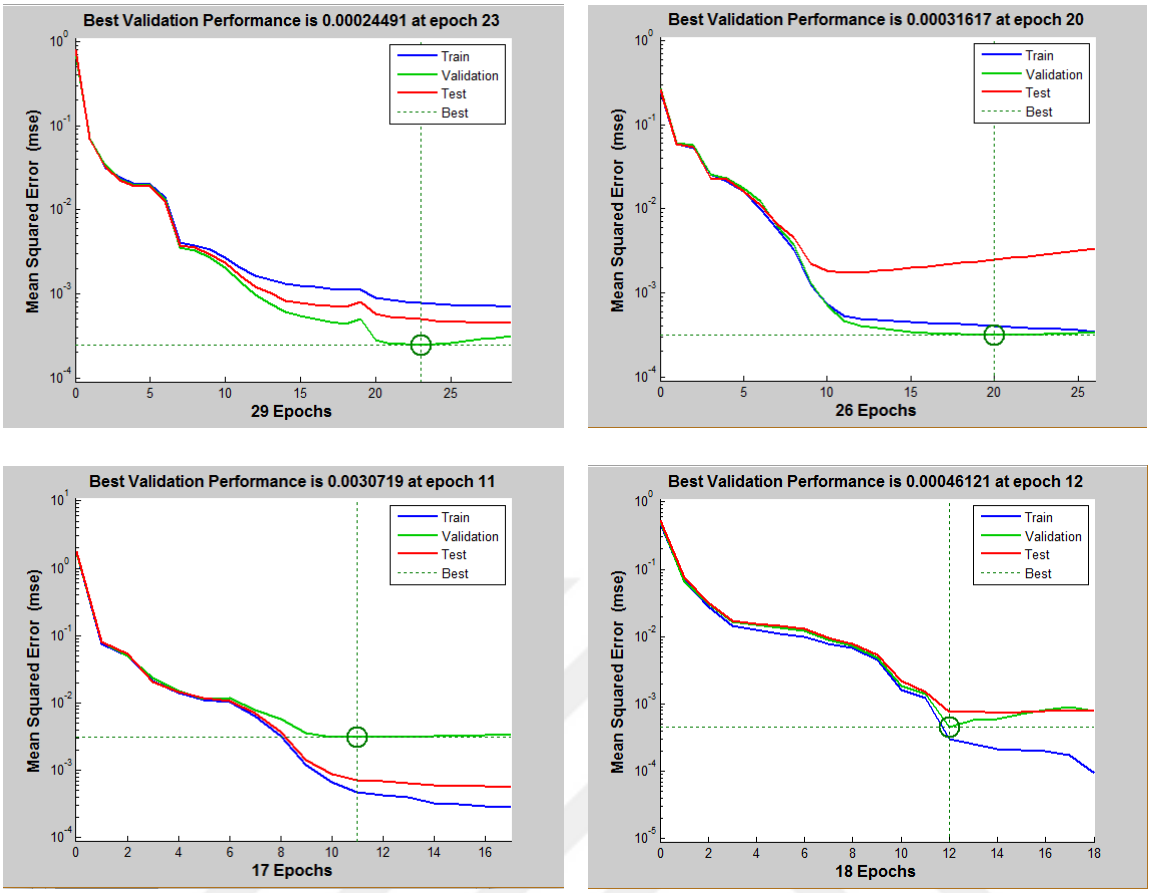
Şekil 5.13. Tek katmanlı modellenen yapay sinir ağı



Şekil 5.14. Çift katmanlı modellenen yapay sinir ağı

Çizelge 5.18. Tek katmanlı yapay sinir ağının farklı sayıda nöronlarda performansı

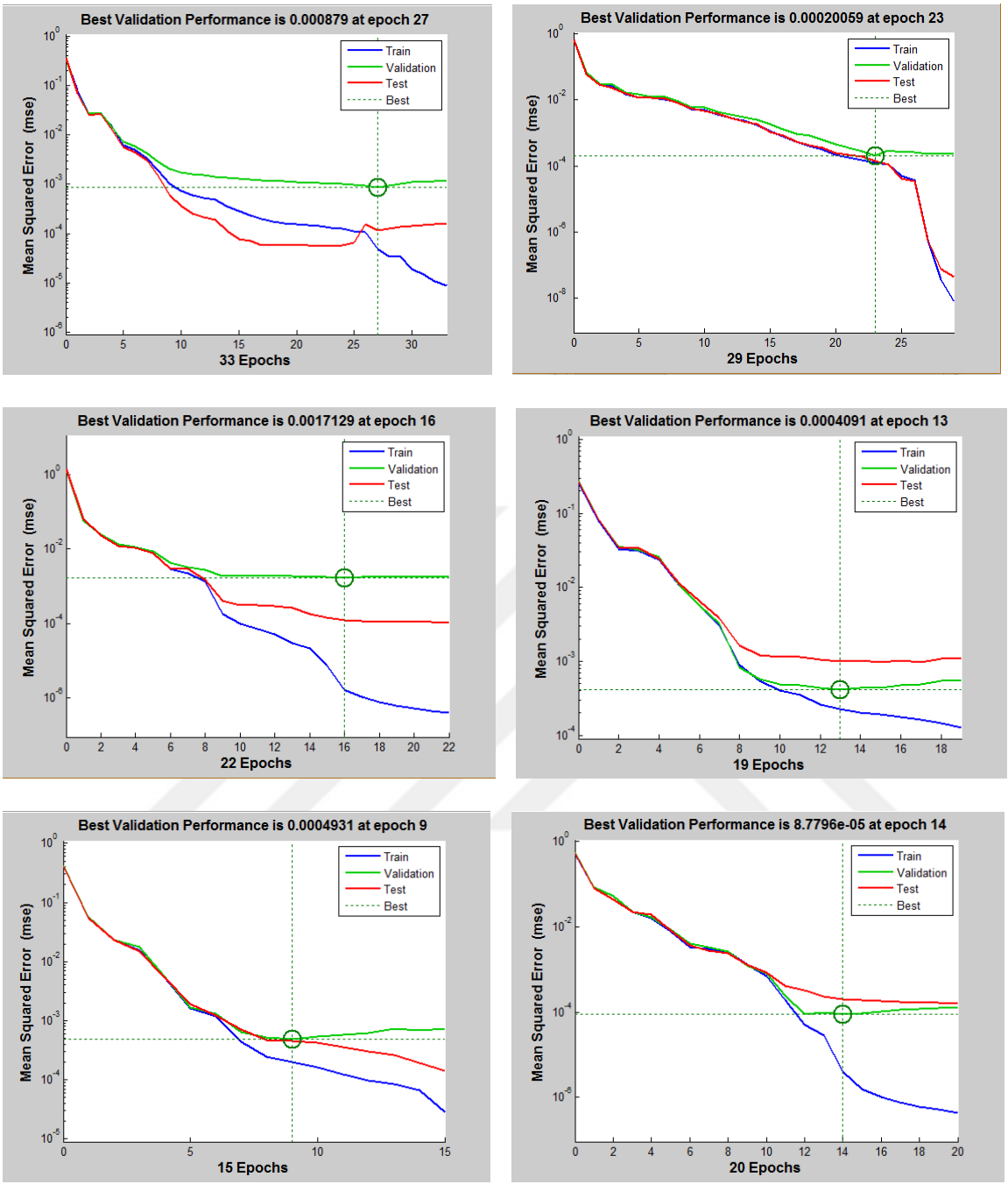
Gizli Katman Sayısı	Nöron Sayısı	Doğruluk Oranı	Hata Oranı	MAPE	R2
1	5	% 89,7153	%10,2847	0,0294	0,949
1	10	%72,6732	%27,3268	0,0997	0,915
1	15	% 60,6885	%39,3115	0,1164	0,948
1	20	%50,9562	%49,0438	0,1656	0,913



Şekil 5.15. Tek katmanlı ve farklı nöron sayılarındaki yapay sinir ağının performansları

Çizelge 5.19. İki katmanlı yapay sinir ağının farklı nöron saylarında performansı

Gizli Katman Sayısı	Birinci Katman Nöron Sayısı	İkinci Katman Nöron Sayısı	Doğruluk Oranı	Hata Oranı	MAPE	R2
2	5	5	% 84,3604	%15,6396	0,0379	0,975
2	5	10	%96,5576	%3,4424	0,0192	0,972
2	10	10	% 93,0302	%6,9698	0,0420	0,978
2	10	15	%87,4203	%12,5797	0,0266	0,980
2	15	15	%95,6226	%4,3774	0,0171	0,991
2	15	20	%87,5478	%12,4522	0,0460	0,984



Şekil 5.16. İki katmanlı ve farklı nöron sayılarındaki yapay sinir ağının performansları

Hem tek katmanda hem de iki katmanda nöron sayıları değiştirilerek deneyler tekrarlanmıştır. Ortalama yüzde hata (MAPE), mutlak değişim yüzdesi (R2) ve başarım oranları performansları değerlendirmek açısından hesaplanmıştır. Elde edilen bu değerler Çizelge 5.18 ve 5.19 da verilmektedir. MAPE değerinin 0'a yaklaşması ve R2 değerinin 1'e yaklaşması performansın iyileştiği anlamına gelmektedir. Tek katmanda ve iki katmanda yapılan deneyler karşılaştırıldığında, iki katmanda yapılan deneylerde MAPE değerleri 0' a daha yakın değerleri alırken, R2 değerleri 1'e daha yakın değerleri almıştır.

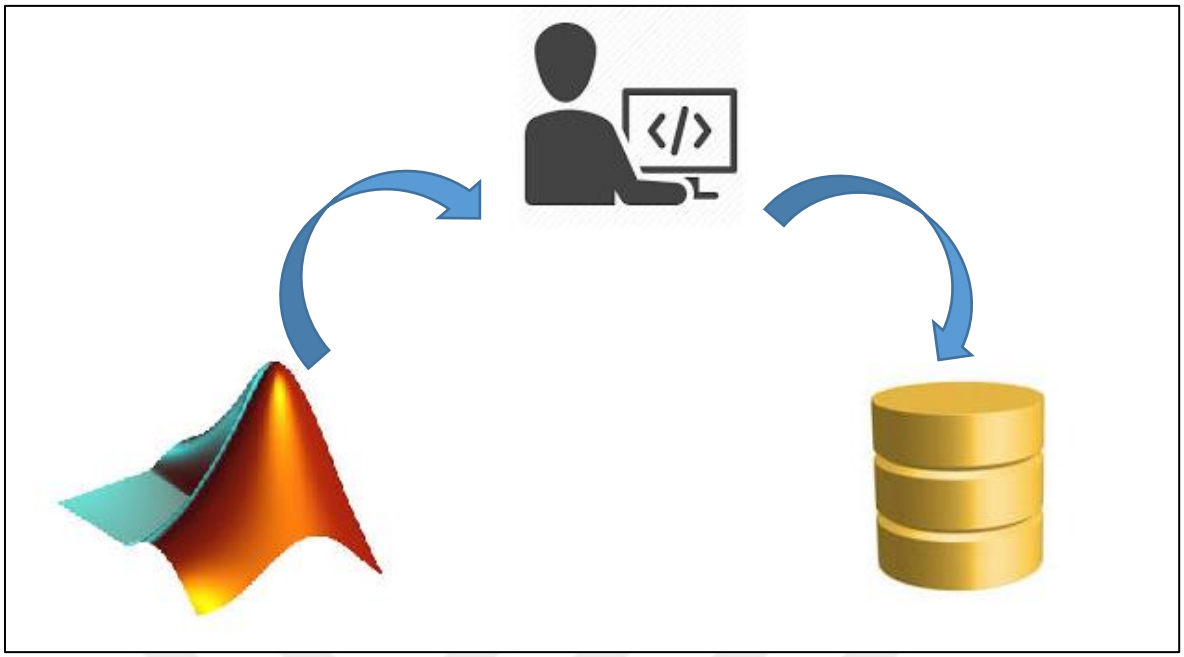
Tek katmanda yapılan analizlere bakıldığında doğruluk oranı nöron sayısının artmasıyla azalmaktadır. Doğruluk oranının en yüksek olduğu nöron sayısı 5'dir. Bu değerde başarım oranı %89 olmaktadır. Yani oluşturulan YSA; 2353 vakadan 2111 tanesini doğru sınıflandırmıştır. Nöron sayısının artması ağın ezberlemesine sebep olmaktadır. Bu durum ağın başarısını olumsuz etkilemekte ve performansın beklenenden düşük çıkmasına neden olmaktadır.

İki katmanda yapılan analizlere bakıldığında; ağın davranışının tek katmandaki gibi olmadığı görülmektedir. Katmanlardaki nöron sayıları arttığında başarım oranının artması söz konusu olmamaktadır. En yüksek başarım oranına %96 ile ikinci analizde ulaşılmıştır. Yani oluşturulan YSA; 2353 vakadan 2272 tanesini doğru sınıflandırmıştır. İkinci analiz ile birlikte ağ, ortalama aynı davranışı göstermiştir.

Genel olarak durum değerlendirildiğinde iki katmanda yapılan deneyler, tek katmana göre daha başarılı olmuştur. Yapay sinir ağlarında katman sayısının artması, öğrenme oranını arttırabilmektedir. Yapılan analizlerde bu durum göze çarpmaktadır. Tek katmanda maksimum %89 olan başarım oranı, iki katmanda %96' ya kadar çıkmıştır.

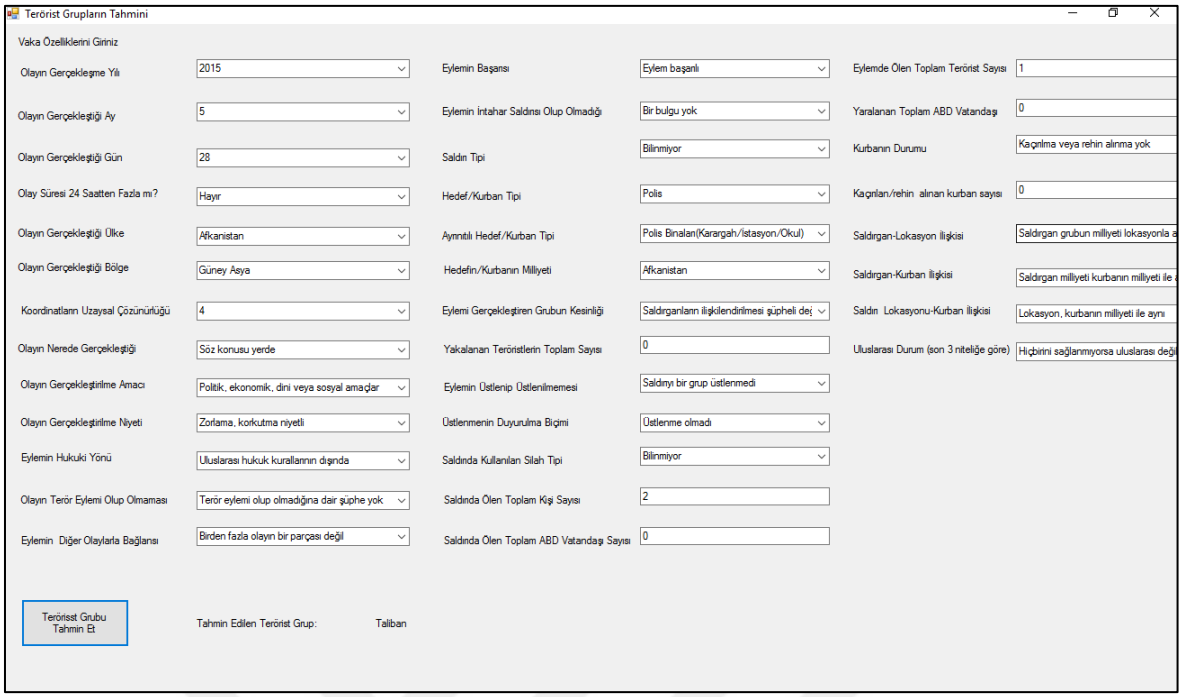
5.2.5. Terörist grubu tahmin sistemi uygulamasının çalıştırılması

Bu kısım uygulamanın kullanıcı arayüz kısmını oluşturmaktadır. Oluşturulan arayüz ile hem veritabanına bağlanıp eğitim ve test kümeleri çekilmekte, hem de Matlab'a bağlanıp Matlab'da yazılan fonksiyon çalıştırılmaktadır. Kullanıcı arayüz kısmı ile diğer birimler arasındaki iletişim şekil 5.17'de gösterilmiştir.



Şekil 5.17. Birimler arası iletişim

Kullanıcı arayüzünde, eğitilen ağa daha önce gösterilmemiş bir vaka verilerek, vakanın hangi terörist gruba ait olduğu tahmin edilmeye çalışılmaktadır. Uygulama çalıştırıldığında kullanıcıdan beklenen; terörist eylemi gerçekleştiren grubun tahmin edilebilmesi için, vakanın özelliklerini girmesidir. Vakanın özellikleri veri tabanından çekilmeyip, kullanıcı tarafından kullanıcı arayüzünden girilmektedir. Vakanın tüm özellikleri girildikten sonra "Terörist Grubu Tahmin Et" butonuna tıklanmaktadır. Daha sonra uygulama, girilen vakayı değerlendirerek sonuç üretmektedir. Örnek üretilen sonuç Şekil 5.18'de verilmiştir.



Şekil 5.18. Terörist grubun tahmin edilmesi

Görüldüğü gibi uygulama girilen vaka özelliklerinden yola çıkarak, “Taliban” terör örgütünü tahmin etmiştir. Bu bağlamda uygulamanın kullanıcı arayüz kısmı, eğitilen ağın test edilme aşamasını oluşturmaktadır.

6. SONUÇ VE ÖNERİLER

Çalışmanın ilk bölümünde; NIBRS kullanılarak, farklı suç türleri üzerinde çeşitli analizler yapılmıştır. İlk analiz suçun bilinmeyen bir unsurunun bilinen unsurları kullanılarak tahmin edilebilmesi olmuştur. Bazen bir suç olayının çözülebilmesi için kilit nokta denilebilecek bazı unsurların bilinmesi gerekebilmektedir. Yani bu unsurun bilinmesi suçun çözülmesi bakımından hayati önem taşıyabilmektedir. Yapılan tahminlerin %90' a varan yüksek doğruluğa ulaşması, suç analizinde veri madenciliği ve makine öğrenmesinin başarısını ortaya koymaktadır. Bu analiz; bir suç vakasının karanlık taraflarının olmasından dolayı çözülememesi sorununa yardımcı olacak bir yaklaşımdır. İkinci analiz daha çok suçu anlamaya yönelik yapılmıştır. Burada amaç; suç içinde direk tahmin edilemeyen, ancak çeşitli analizler yapılınca ortaya çıkarılabilecek, ilişki desenlerinin ortaya çıkarılmasıdır. Birliktelik kuralları modeli ve bu modeli uygulamak için kullanılan apriopri algoritması ile kurallar keşfedilmiştir. Bu birliktelik kuralları yine suçun çözümünde bilinmeyen unsurların ortaya çıkarılmasını sağlayacak ve bu doğrultuda suçların çözülmesine olanak sağlayacaktır. Son analiz ise suç türünün anlaşılmasına yönelik olarak yapılmıştır. Her suç kendine özgü bir desene sahiptir. Yani suçlu profili olabileceği gibi, suç profili de bulunmaktadır. Burada amaç çok sayıda vaka içerisinden aynı suç türündeki vakaların sınıflara ayrılabilmesidir. Yapay sinir ağı algoritmasının veri kümesi üzerine uygulanmasıyla %97 gibi yüksek bir başarım oranı elde edilmiştir. Doğruluk oranını iyileştirmeye yönelik olarak, yapay sinir ağı üzerine uygulanan önyargılı örneklem dağılımı yaklaşımı ile de başarım oranı %98 olmuştur. Bu doğrultuda; uygulanan yöntem, başarı yüzdesi olarak suç analizinde kullanılabilecek bir yaklaşımdır.

Çalışmanın ikinci kısmında; terörist grupların gerçekleştirdiği eylemlerin ayrıntılı bilgisinden oluşan GTD veritabanı kullanılarak terörist grubu tahmin sistemi oluşturulmuştur. Bu sistem; suçlu tespiti amacıyla tasarlanmıştır. Gerçekleştirilen yazılımda; veri kümesinin eğitilmesi için yapay sinir ağları kullanılmıştır. Tek katmanda ve iki katmanda nöron sayıları değiştirilerek, öğrenme test edilmiştir. Tek katmanda başarım %89,71'de kalırken, iki katmanda %96,5576'ya kadar ulaşılmıştır. Elde edilen başarım oranlarına bakıldığında, gerçekleşen eylemlerin faillerinin yakalanması ya da gelecekte gerçekleştirilebilecek eylemlerin tespit edilebilmesi bakımından bu sistem faydalı sonuçlar üretebilmektedir.

Genel olarak yapılan tez çalışmasında amaç; suçun çözümünde güvenlik güçlerine farklı bakış açıları sunabilmektedir. Ülkemizde maalesef güvenlik güçleri geleneksel yöntemleri kullanarak suçları çözmeye, suçluları yakalamaya çalışmaktadır. Gelecekte oluşabilecek suçlar için ise, bilinçlendirme sunumları yapmak dışında, yapabilecekleri çok fazla şey bulunmamaktadır. Veri madenciliği ve makine öğrenmesi ya da genel olarak yapay zeka teknolojileri kullanılarak oluşturulacak sistemler, suçun çözülmesi veya gelecekte meydana gelebilecek suçların önlenmesi bakımından güvenlik güçlerine faydalı olacaktır.

KAYNAKLAR

1. Öztemel E. (2003). *Yapay Sinir Ağları*. (Üçüncü Baskı) İstanbul: Papatya Yayıncılık.

2. Özkan, Y. (2013). *Veri madenciliği yöntemleri* (İkinci Baskı). İstanbul: Papatya Yayıncılık.

3. Sokullu-Akıncı, F. (2014). *Kriminoloji* (Onbirinci Baskı). İstanbul: Beta Yayınları.

4. Yu, C., Ward, M. W., Morabito, M., Ding, W. (2011). Crime Forecasting Using Data Mining Techniques. Paper presented at the *11th IEEE International Conference on Data Mining Workshops*, 779-786.

5. Brown, D.E. (1998). *The Regional Crime Analysis Program (RECAP): A framework for Mining Data to Catch Criminals*. Paper presented at the Systems, Man, and Cybernetics, 1998. 1998 IEEE International Conference on , 2848-2853.

6. Takçı, H., Hayta, Ş. (2014). *Suç Veri Madenciliği Yardımıyla Hırsızlık Suçları Hakkında Kural Çıkarımı*. Eleco Elektrik – Elektronik – Bilgisayar ve Biyomedikal Mühendisliği Sempozyumunda sunuldu, 694-699.

7. Bruin, J.S., Cocx, T.K., Kosters, W. A., Laros, J., Kok, J.N. (2006). *Data mining approaches to criminal career analysis*. Paper presented at the Sixth International Conference on Data Mining (ICDM') (ICDM'06), 171-177.

8. Nath,S.V. (2006). *Crime Pattern Detection Using Data Mining*. Paper presented at the International Conferences on Web Intelligence and Intelligent Agent Technology - Workshops, 41-44.

9. Saeed, U., Sarim, M., Usmani, A., Mukhtar, A., Shaikh, A., B., Raffat, S.,K. (2015). *Application of Machine Learning Algorithims in Crime Classification and Classification Rule Mining*. Research Journal of Recent Sciences, Vol. 4(3), 106-114.

10. Almanie, T., Mirza, R., & Lor, E. (2015). *Crime Prediction Based On Crime Types And Using Spatial And Temporal Criminal Hotspots*. International Journal of Data Mining & Knowledge Management Process, 5(4), 10.5121/ijdkp.2015.5401.

11. Yang, R., & Olafsson, S. (2011). Classification for predicting offender affiliation with murder victims. *Expert Systems with Applications*, 38(11), 13518-13526.

12. Asmai, S. A., Roslin, N. I. A., Abdullah, R. W., & Ahmad, S. (2014). *Predictive crime mapping model using association rule mining for crime analysis.* Paper presented at the International Semposium on Research in Innovation and Sustainability. 26(5), 1703-1706.

13. Arulanandam, R., Savarimuthu, B. T. R., & Purvis, M. A. (2014). *Extracting crime information from online newspaper articles*. Paper presented at the Second Australasian Web Conference. 155.

14. Baumgartner, K. C., Ferrari, S., & Salfati, C. G. (2005). *Bayesian network modeling of offender behavior for criminal profiling.* Paper presented at the Decision and Control, 2005 and 2005 European Control Conference. *CDC-ECC'05.*

15. Iqbal, R., Murad, M. A. A., Mustapha, A., Panahy, P. H. S., & Khanahmadliravi, N. (2013). An experimental study of classification algorithms for crime prediction. *Indian Journal of Science and Technology*, 6(3), 4219-4225.

16. Sathyadevan, S., Devan, M., & Surya Gangadharan, S. (2014). Crime analysis and prediction using data mining. Paper presented at the *Networks & Soft Computing (ICNSC), 2014 First International Conference on*.

17. Encylopaedia Brittanica (Vol. 1). Edinburgh: A. Bell and C. Macfarquhar.

18. İnternet: Webster's Online Dictionary, URL: http://www.webstersonline-dictionary.org. Son Erişim Tarihi: 03.12.2016.

19. Polat, O. (2009). *Kriminoloji ve Kriminalistik Üzerine Notlar*. Ankara: Seçkin Yayınları.

20. Champion, D. J. (2005). *The American Dictionary of Criminal Justice*. Los Angeles.

21. Dolu O. (2012). *Suç Teorileri* (4. Baskı). Ankara: Seçkin Yayınları.

22. Demirbaş T. (2014). Kriminoloji (5. Baskı). Ankara: Seçkin Yayıncılık.

23. Kalıkov, A. (2006). *Veri Madenciliği ve Bir E-Ticaret Uygulaması*. Yüksek Lisans Tezi, Gazi Üniversitesi Fen Bilimleri Enstitüsü, Ankara.

24. İnan, O. (2003). *Veri Madenciliği*. Yüksek Lisans Tezi, Selçuk Üniversitesi Fen Bilimleri Enstitüsü, Konya.

25. Hand, D. J. (1998). Data mining: Statistics and more? *The American Statistician*, 52(2), 112-118.

26. Jacobs, P. (1999). Data mining: what general managers need to know. *Harvard Management Update*, 4(10), 8.

27. Doğan, Ş. Ve Türkoğlu, İ. (2007). Hypothyroidi and Hyperthyroidi Detection from Thyroid Hormone Parameters by Using Decision Trees. *Doğu Anadolu Bölgesi Araştırmaları Dergisi*, 5(2), 163-169.

28. Kittler, R. and Wang, W. (1999). The emerging role for data mining. *Solid state technology*. 42(11), 45-45.

29. Ünsal, Ö. (2011). *Mesleki Alan Seçimlerinin Makine Öğrenmesi Algoritması Kullanılarak Belirlenmesi*. Yüksek Lisans Tezi, Gazi Üniversitesi Bilişim Enstitüsü, Ankara.

30. Davis B. (1999). Data Mining Transformed. *Information Week* , 751: 86.

31. Silahtaroğlu, G. (2013). *Veri Madenciliği Kavram ve Algoritmaları* (2. Baskı). İstanbul: Papatya Yayıncılık.

32. Han, J., & Kamber, M. (2001). Data mining: concepts and techniques. *Morgan Kauffman.*

33. İnternet: Veri Madenciliği. (2013). URL: http://ab.org.tr/ab13/bildiri/175.pdf. Son Erişim Tarihi: 02.11.2016.

34. Dunham, M.H. (2003). Data Mining Introductory and Advanced Topics. *Prentice Hall*, *Pearson Education, New Jersey.*

35. Özdamar, E., O. (2002). *Veri Madenciliğinde Kullanılan Teknikler ve Bir Uygulama*. Yüksek Lisans Tezi, Mimar Sinan Üniversitesi Fen Bilimleri Enstitüsü, İstanbul.

36. İnternet: Orhon, U. Makine öğrenmesi. URL: , http://bmb.cu.edu.tr/uorhan/DersNotu/Ders01.pdf.Son Erişim Tarihi: 02.11.2016.

37. Topuz, M. D. (2014). *Makine Öğrenmesi Algoritmaları ve Anomali Tespiti*. Yüksek Lisans Tezi, Bahçeşehir Üniversitesi Fen Bilimleri Enstitüsü, İstanbul.

38. Dalyan, T. (2006). *Makine öğrenmesinde 1R algoritması ve ikinci kuralın(2R) oluşturulması*. Yüksek Lisans Tezi, Kocaeli Üniversitesi Fen Bilimleri Enstitüsü, Kocaeli.

39. Maindonald, J., & Braun, J. (2006). *Data analysis and graphics using R: an example-based approach* (Vol. 10). Cambridge University Press.

40. İnternet: Şeker, Ş.E. (2008). SVM (Support Vector Machine, Destekçi Vektör Makinesi). URL: http://www.webcitation.org/query?url=http%3A%2F%2Fbilgisayarkavramlari.sadievrenseker.com%2F2012%2F01%2F29%2Fistatistiksel-normallestirme-statistical-normalisation%2F. Son Erişim Tarihi: 02.11.2016.

41. Vapnik, V. N., & Vapnik, V. (1998). *Statistical learning theory* (Vol. 2). Wiley New York.

42. Eray, O. (2008). *Destek vektör makineleri ile ses tanıma uygulaması*. Yüksek Lisans Tezi, Pamukkale Üniversitesi Fen Bilimleri Enstitüsü, Denizli.

43. Shen, J., Pei, Z., Fisher, G., & Lee, E. (2006). Modelling and analysis of waviness reduction in soft-pad grinding of wire-sawn silicon wafers by support vector regression. *International journal of production research*, 44(13), 2605-2623.

44. Meral, M., & Diri, B. (2014). *Sentiment analysis on Twitter*. Paper presented at the Signal Processing and Communications Applications Conference (SIU), 2165-0608.

45. Witten, I. H., & Frank, E. (2005). *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann.

46. Sayed, S. (2011). *Real Time Data Mining*. Canada: Self Help Publishers.

47. Gülce, A. C. (2010). *Veri madenciliğinde apriori algoritması ve apripori algoritmasının farklı veri kümelerinde uygulanması*. Yüksek Lisans Tezi, Trakya Üniversitesi Fen Bilimleri Enstitüsü, Edirne.

48. Agrawal, R., & Srikant, R. (1994). *Fast algorithms for mining association rules*. Paper presented at the Proc. 20th international conferance very large data bases, *VLDB*.

49. Bosque, M. (2004). *Web-based Neural Nets: Interactive Artificial Neural Networks For The Internet: Introduction And Tricks*. iUniverse.

50. Bishop, C.M. (2006). *Pattern Recognition and Machine Learning.* Springer.

51. Srinivasan, K., Fisher, D. (1995). Machine learning approaches to estimating software development effort. *IEEE Transactions on Software Engineering*, 21(2), 126-137.

52. İnternet: Çayıroğlu, İ. Yapay Sinir Ağları. URL: http://www.ibrahimcayiroglu.com/dokumanlar/ilerialgoritmaanalizi/ilerialgoritmaanalizi-5.hafta-yapaysiniraglari.pdf Son Erişim Tarihi: 03.11.2016.

53. Bakırlı, G. (2016). *Machine Learning Approach to Create Intelligent Decision-Maker System*. Doktora Tezi, Dokuz Eylül Üniversitesi, Fen Bilimleri Enstitüsü, İzmir.

54. İnternet: Yalçın, N. Yapay Sinir Ağları. URL: http://musaatas.siirt.edu.tr/ANN/YSA_Giris.pdf. Son Erişim Tarihi: 03.11.2016.

55. İnternet: National Incident-Based Reporting System Resource Guide. URL: http://www.icpsr.umich.edu/icpsrweb/NACJD/NIBRS/ .Son Erişim Tarihi: 05.10.2016.

56. İnternet: National Incident-Based Reporting System (NIBRS) Series – icpsr. URL: https://www.icpsr.umich.edu/icpsrweb/NACJD/series/128/studies?archive=NACJD&sortBy=7 .Son Erişim Tarihi: 05.10.2016.

57. İnternet: Global Terorrism Database (GTD). URL: https://www.start.umd.edu/gtd/. Son Erişim Tarihi: 05.10.2016.

58. İnternet: Global Terorrism Database Codebook: Inclusion criteria and variables. (2016). URL: https://www.start.umd.edu/gtd/downloads/Codebook.pdf .Son Erişim Tarihi: 05.10.2016.

59. İnternet: Weka 3: Data Mining Software in Java. URL: http://www.cs.waikato.ac.nz/ml/weka/. Son Erişim Tarihi: 02.12.2016.

60. İnternet: MATLAB. URL: https://www.mathworks.com/products/matlab.html. Son Erişim Tarihi: 02.12.2016.

61. İnternet: National Incident-Based Reporting System, 2005 (ICPSR 4720). URL: https://www.icpsr.umich.edu/icpsrweb/NACJD/series/128/studies/4720?archive=NACJD&sortBy=7 .Son Erişim Tarihi: 05.10.2016.

62. İnternet: Uniform Crime Reporting Program Data: National Incident-Based Reporting System 2013 (ICPSR 36120). URL: https://www.icpsr.umich.edu/icpsrweb/NACJD/series/128/studies/36120?archive=NACJD&sortBy=7 .Son Erişim Tarihi: 05.10.2016.

EKLER

EK-1. NIBRS'de bulunan nitelikler

Kod	Nitelik
	V1000s – İdari (Administrative)
V1006	REPORT DATE INDICATOR
V1007	INCIDENT DATE HOUR
V1008	TOTAL OFFENSE SEGMENTS
V1009	TOTAL VICTIM SEGMENTS
V1010	TOTAL OFFENDER SEGMENTS
V1011	TOTAL ARRESTEE SEGMENTS
V1012	CITY SUBMISSION
V1013	CLEARED EXCEPTIONALLY
V1014	EXCEPTIONAL CLEARANCE DATE
V1016	CARGO THEFT
	V2000s –Suç (Offense)
V20061- V20063	UCR OFFENSE CODE
V20071- V20073	OFFENSE ATTEMPTED/COMPLETED
V20081- V20083	OFFENDER(S) SUSPECTED OF USING
V20091- V20103	OFFENDER(S)OF USING
V20111- V20113	LOCATION TYPE
V20121- V20123	NUMBER OF PREMISES ENTERED
V20131- V20133	METHOD OF ENTRY
V20141- V20163	TYPE OF CRIMINAL ACTIVITY
V20171- V20193	TYPE WEAPON/FORCE INVOLVED 1
V20201- V20203	BIAS MOTIVATION
	V3000s – Eşya (Property)
V30061- V30063	TYPE PROPERTY LOSS/ETC
V30071- V30073	PROPERTY DESCRIPTION
V30081- V30083	VALUE OF PROPERTY
V30091- V30093	DATE RECOVERED
V30101- V30103	NUMBER OF STOLEN MOTOR VEHICLES
V30111- V30113	NUMBER OF RECOVERED MOTOR VEHICLES
V30121- V30123	SUSPECTED DRUG TYPE
V30131- V30133	ESTIMATED QUANTITY
V30141- V30143	ESTIMATED QUANTITY - FRACTIONAL THOUSANDTHS
V30151- V30233	TYPE MEASUREMENT
V30161- V30203	SUSPECTED DRUG TYPE
V30171- V30213	ESTIMATED QUANTITY
V30181- V30223	ESTIMATED QUANTITY - FRACTIONAL THOUSANDTHS
	V4000s – Kurban (Victim)
V4006	VICTIM SEQUENCE NUMBER
V4007- V4016	UCR OFFENSE CODE
V4017	TYPE OF VICTIM
V4017A	TYPE OF ACTIVITY

EK-1. (devam) NIBRS'de bulunan nitelikler

Kod	**Nitelik**
V4017B	ASSIGNMENT TYPE
V4017C	ORI OTHER JURISDICTION
V4018	AGE OF VICTIM
V4019	SEX OF VICTIM
V4020	RACE OF VICTIM
V4021	ETHNICITY OF VICTIM
V4022	RESIDENT STATUS OF VICTIM
V4023- V4024	AGGRAVATED ASSAULT/HOMICIDE CIRCUMSTANCES
V4026- V4030	TYPE INJURY
V4031- V4049	OFFENDER NUMBER TO BE RELATED
V4032- V4050	RELATIONSHIP OF VICTIM TO OFFENDER
V40061- V40063	VICTIM SEQUENCE NUMBER
V40071- V40163	UCR OFFENSE CODE
V40171- V40173	TYPE OF VICTIM
V4017A1- V4017A3	TYPE OF ACTIVITY
V4017B1- V4017B3	ASSIGNMENT TYPE
V4017C1- V4017C3	ORI OTHER JURISDICTION
V40181- V40183	AGE OF VICTIM
V4017B1- V4017B3	ASSIGNMENT TYPE
V4017C1- V4017C3	ORI OTHER JURISDICTION
V40181- V40183	AGE OF VICTIM
V4019- V40193	SEX OF VICTIM
V40201- V40203	RACE OF VICTIM
V40213- V40213	ETHNICITY OF VICTIM
V40221- V40223	RESIDENT STATUS OF VICTIM
V40231- V40243	AGGRAVATED ASSAULT/HOMICIDE CIRCUMSTANCES
V40251- V40253	ADDITIONAL JUSTIFIABLE HOMICIDE CIRCUMSTANCES
V40261- V40303	TYPE INJURY
V40311- V40493	OFFENDER NUMBER TO BE RELATED
V40321- V40503	RELATIONSHIP OF VICTIM TO OFFENDER
	V5000s – Suçlu (Offender)
V5006	OFFENDER SEQUENCE NUMBER
V5007	AGE OF OFFENDER
V5008	SEX OF OFFENDER
V5009	RACE OF OFFENDER
V5011	ETHNICITY OF OFFENDER
V50061- V50063	OFFENDER SEQUENCE NUMBER
V50071- V50073	AGE OF OFFENDER
V50081- V50083	SEX OF OFFENDER

EK-1. (devam) NIBRS'de bulunan nitelikler

Kod	**Nitelik**
V50091- V50093	RACE OF OFFENDER
V50111- V50113	ETHNICITY OF OFFENDER
	V6000s – Tutuklu (Arrestee)
V6006	ARRESTEE SEQUENCE NUMBER
V6007	ARREST TRANSACTION NUMBER
V6008	ARREST DATE
V6009	TYPE OF ARREST
V6010	MULTIPLE ARRESTEE SEGMENTS INDICATOR
V6011	UCR ARREST OFFENSE CODE
V6012- V6013	ARRESTEE ARMED WITH
V6014	AGE OF ARRESTEE
V6015	SEX OF ARRESTEE
V6016	RACE OF ARRESTEE
V6017	ETHNICITY OF ARRESTEE
V6018	RESIDENT STATUS OF ARRESTEE
V6019	DISPOSITION OF ARRESTEE UNDER 18
V60061- V60063	ARRESTEE SEQUENCE NUMBER
V60071- V60073	ARREST TRANSACTION NUMBER
V60081- V60083	ARREST DATE
V60091- V60093	TYPE OF ARREST
V60101- V60103	MULTIPLE ARRESTEE SEGMENTS INDICATOR
V60111- V60113	UCR ARREST OFFENSE CODE
V60121- V60133	ARRESTEE ARMED WITH
V60141- V60143	AGE OF ARRESTEE
V60151- V60153	SEX OF ARRESTEE
V60161- V60163	RACE OF ARRESTEE
V60171- V60173	ETHNICITY OF ARRESTEE
V60181- V60183	RESIDENT STATUS OF ARRESTEE
V60191- V60193	DISPOSITION OF ARRESTEE UNDER 18
	All_Offenses - Concatenated All Offenses Variable
ALLOFNS	ALL OFFENSE CODES FOR THE INCIDENT

EK-2. GTD'de bulunan nitelikler

Kod	Nitelik
	GTD ID and Date
eventid	GTD ID
iyear	YEAR
imonth	MONTH
iday	DAY
approxdate	APPROXİMATE DATE
extended	EXTENDED İNCİDENT
resolution	DATE OF EXTENDED İNCİDENT RESOLUTİON
	Incident Information
summary	INCIDENT SUMMARY
crit1	POLITICAL, ECONOMIC, RELIGIOUS, OR SOCIAL GOAL (CRIT2)
crit2	INTENTION TO COERCE, INTIMIDATE OR PUBLICIZE TO LARGER AUDIENCE(S) (CRIT2)
crit3	OUTSIDE INTERNATIONAL HUMANITARIAN LAW (CRIT3)
doubtterr	DOUBT TERRORİSM PROPER?
alternative	ALTERNATİVE DESIGNATION
alternative_txt	ALTERNATİVE DESIGNATION
multiple	PART OF MULTİPLE INCIDENT
related	RELATED INCIDENT
	Incident Location
country	COUNTRY
country_txt	COUNTRY
region	REGION
region_txt	REGION
provstate	PROVINCE/ ADMINISTRATIVE REGION/ STATE
city	CITY
vicinity	VICINITY
location	LOCATION DESCRIPTION
latitude	LATITUDE
longitude	LONGITUDE
specificity	GEOCODING SPECIFICITY
	Attack Information
attacktype1	ATTACK TYPE
attacktype1_txt	ATTACK TYPE
attacktype2	SECOND ATTACK TYPE
attacktype2_txt	SECOND ATTACK TYPE
attacktype3	THİRD ATTACK TYPE
attacktype3_txt	THİRD ATTACK TYPE
success	SUCCESSFUL ATTACK
suicide	SUICIDE ATTACK

EK-2. (devam) GTD'de bulunan nitelikler

Kod	Nitelik
	Weapon Information
weaptype1	WEAPON TYPE
weaptype1_txt	WEAPON TYPE
weapsubtype1	WEAPON SUB-TYPE
weapsubtype1_txt	WEAPON SUB-TYPE
weaptype2	SECOND WEAPON TYPE
weaptype2_txt	SECOND WEAPON TYPE
weapsubtype2	SECOND WEAPON SUB-TYPE
weapsubtype2_txt	SECOND WEAPON SUB-TYPE
weaptype3	THİRD WEAPON TYPE
weaptype3_txt	THİRD WEAPON TYPE
weapsubtype3	THİRD WEAPON SUB-TYPE
weapsubtype3_txt	THİRD WEAPON SUB-TYPE
weaptype4	FOURTH WEAPON TYPE
weaptype4_txt	FOURTH WEAPON TYPE
weapsubtype4	FOURTH WEAPON SUB-TYPE
weapsubtype4_txt	FOURTH WEAPON SUB-TYPE
weapdetail	WEAPON DETAİLS
	Target/Victim Information
targtype1	TARGET/ VICTIM TYPE
targtype1_txt	TARGET/ VICTIM TYPE
targsubtype1	TARGET/ VICTIM SUBTYPE
targsubtype1_txt	TARGET/ VICTIM SUBTYPE
corp1	NAME OF ENTITY
target1	SPECIFIC TARGET/ VICTIM
natlty1	NATIONALITY OF TARGET/VICTIM
natlty1_txt	NATIONALITY OF TARGET/VICTIM
targtype2	SECOND TARGET/ VICTIM TYPE
targtype2_txt	SECOND TARGET/ VICTIM TYPE
targsubtype2	SECOND TARGET/ VICTIM SUBTYPE
targsubtype2_txt	SECOND TARGET/ VICTIM SUBTYPE
corp2	NAME OF SECOND ENTITY
target2	SECOND SPECIFIC TARGET/ VICTIM
natlty2	NATIONALITY OF SECOND TARGET/VICTIM
natlty2_txt	NATIONALITY OF SECOND TARGET/VICTIM
targtype3	THİRD TARGET/ VICTIM TYPE
targtype3_txt	THİRD TARGET/ VICTIM TYPE
targsubtype3	THİRD TARGET/ VICTIM SUBTYPE
targsubtype3_txt	THİRD TARGET/ VICTIM SUBTYPE
corp3	NAME OF THİRD ENTITY
target3	THİRD SPECIFIC TARGET/ VICTIM

EK-2. (devam) GTD'de bulunan nitelikler

Kod	Nitelik
natlty3	NATIONALITY OF THİRD TARGET/VICTIM
natlty3_txt	NATIONALITY OF THİRD TARGET/VICTIM
	Perpetrator Information
gname	PERPATRATOR GROUP NAME
gsubname	PERPATRATOR SUB-GROUP NAME
gname2	SECOND PERPATRATOR GROUP NAME
gsubname2	SECOND PERPATRATOR SUB-GROUP NAME
gname3	THİRD PERPATRATOR GROUP NAME
gsubname3	THİRD PERPATRATOR SUB-GROUP NAME
guncertain1	FİRST PERPATRATOR GROUP SUSPECTED/ UNCORFİRMED ?
guncertain2	SECOND PERPATRATOR GROUP SUSPECTED/ UNCORFİRMED ?
guncertain3	THİRD PERPATRATOR GROUP SUSPECTED/ UNCORFİRMED ?
nperps	NUMBER OF PERPATRATORS
nperpcap	NUMBER OF PERPATRATORS CAPTURED
claimed	CLAIM OF RESPONSIBILITY?
claimmode	MODE FOR CLAIM OF RESPONSIBILITY
claimmode_txt	MODE FOR CLAIM OF RESPONSIBILITY
compclaim	COMPETING CLAIMS OF RESPONSIBILITY?
claim2	SECOND GROUP CLAIM OF RESPONSIBILITY?
claimmode2	MODE FOR CLAIM OF RESPONSIBILITY
claim3	THİRD GROUP CLAIMS OF RESPONSIBILITY?
claimmode3	MODE FOR THİRD GROUP CLAIM OF RESPONSIBILITY
motive	MOTIVE
	Casualties and Consequences
nkill	TOTAL NUMBER OF FATALITIES
nkillus	NUMBER OF US FATALITIES
nkillter	NUMBER OF PERPETRATOR FATALITIES
nwound	TOTAL NUMBER OF INJURED
nwoundus	NUMBER OF US INJURED
nwoundte	NUMBER OF PERPETRATOR INJURED
property	PROPERTY DAMAGE
propextent	EXTEND OF PROPERTY DAMAGE
propextent_txt	EXTEND OF PROPERTY DAMAGE
propvalue	VALUE OF PROPERTY DAMAGE (IN USD)
propcomment	PROPERTY DAMAGE COMMENTS
ishostkid	HOSTAGES OR KIDNAPPING VICTIMS
nhostkid	TOTAL NUMBER OF HOSTAGES/ KIDNAPPING VICTIMS
nhostkidus	NUMBER OF U.S. HOSTAGES/ KIDNAPPING VICTIMS
nhours	HOURS OF KIDNAPPING/ HOSTAGE INCIDENT
ndays	DAYS OF KIDNAPPING/ HOSTAGE INCIDENT

EK-2. (devam) GTD'de bulunan nitelikler

Kod	Nitelik
divert	COUNTRY THAT KIDNAPPERS/ HIJACKING DIVERTED TO
kidhijcountry	COUNTRY OF KIDNAPPERS/ HIJACKING RESOLUTION
ransom	RANSOM DEMANDED
ransomamt	TOTAL RANSOM AMOUNT DEMANDED
ransomus	RANSOM DEMANDED FROM U.S. SOURCE (S)
ransomamtus	RANSOM AMOUNT DEMANDED FROM U.S. SOURCE (S)
ransompaid	TOTAL RANSOM AMOUNT PAID
ransompaidus	RANSOM AMOUNT PAID BY U.S. SOURCES
ransomnote	RANSOM NOTES
hostkidoutcome	KIDNAPPING/ HOSTAGE OUTCOME
hostkidoutcome_txt	KIDNAPPING/ HOSTAGE OUTCOME
nreleased	NUMBER RELEASED/ESCAPED/RESCUED
	Additional Information and Sources
addnotes	ADDITIONAL NOTES
INT_LOG	INTERNATIONAL -LOGISTICAL
INT_IDEO	INTERNATIONAL -IDEOLOGICAL
INT_MISC	INTERNATIONAL -MISCELLANEOUS
INT_ANY	INTERNATIONAL –ANY OF THE ABOVE
scite1	FIRST SOURCE CITATION
scite2	SECOND SOURCE CITATION
scite3	THİRD SOURCE CITATION
dbsource	DATA COLLECTION

ÖZGEÇMİŞ

Kişisel Bilgiler

Soyadı, adı : ORAKCI, Merve

Uyruğu : T.C

Doğum tarihi ve yeri : 22/11/1991 Sivas

Medeni hali : Bekar

Telefon : 0 (312) 202 38 14

e-posta : merveorakci@gazi.edu.tr

Eğitim Derecesi	**Okul/Program**	**Mezuniyet yılı**
Yüksek lisans	Gazi Üniversitesi / Adli Bilişim A.B.D	Devam Ediyor
Lisans	Eskişehir Osmangazi Üniversitesi / Elektrik Elektronik Mühendisliği	2014
Lisans	Eskişehir Osmangazi Üniversitesi / Bilgisayar Mühendisliği Bölümü	2014
Ön Lisans	Anadolu Üniversitesi / Açıköğretim Fakültesi / Adalet	Devam Ediyor
Lise	Sivas Anadolu Lisesi	2009

İş Deneyimi, Yıl	**Çalıştığı Yer**	**Görev**
2015-Devam Ediyor	Gazi Üniversitesi	Araştırma Görevlisi

Yabancı Dili

İngilizce

Yayınlar

1. Orakcı, M., Kök, İ., Çakır, H., (2016). Adli Bilişim Eğitiminin Gerkesinimi ve Genel Olarak Değerlendirilmesi. Bilişim Teknolojileri Dergisi / International Journal of Informatic Technologies. 9(2).

2. Orakcı, M., Ciylan, B. (2016). Prediction of Unknown Elements in Rape Cases: Machine Learning Approach in Crime Analysis. International Conference on Research in Education & Science. Mayıs, 2016. Bodrum.

3. Orakcı, M., Ciylan, B. (2016). Exploration of Secret Relations in Homicide Cases. International Conference on Research in Education & Science. Mayıs, 2016. Bodrum.

4. Orakcı, M. (2016). A General View of Computer Forensics Education in Turkey: Exigence of Computer Forensics and its Education. International Conference on Education in Mathematics, Science & Technology. Mayıs, 2016. Bodrum.

5. Orakcı, M., Ciylan, B. (2016). Estimating Victim Type in Credit Card Fraud with the Approach of Artificial Neural Networks. Global Conference on Applied Computing in Science and Engineering. Temmuz, 2016. Roma.

6. Orakcı, M., Ciylan, B., Kök, İ., Sevri, M. (2016). Suç Analizinde Veri Madenciliği Teknikleri ve Makine Öğrenmesi Algoritmalarının Kullanılması. Akademik Bilişim Konferansı. Şubat, 2016. Aydın.

GAZİ GELECEKTİR..