**T.C.**
**GİRESUN ÜNİVERSİTESİ**
**FEN BİLİMLERİ ENSTİTÜSÜ**

# ÇOCUK İSTİSMARI DAVALARININ SONUÇLARINA ETKİ EDEN FAKTÖRLERİN BELİRLENMESİ İÇİN MAKİNE ÖĞRENMESİ YAKLAŞIMLARI

**DOKTORA TEZİ**

**Öğrencinin Adı SOYADI : Saime Şule AKSAKAL**

**Tezin Enstitüye Verildiği Tarih :**

**Enstitü Anabilim Dalı : İstatistik**

**Tez Danışmanı : Prof. Dr. Erol EĞRİOĞLU**

**Mayıs 2025**
**GİRESUN**

**T.C.**
**GİRESUN ÜNİVERSİTESİ**
**FEN BİLİMLERİ ENSTİTÜSÜ**

# ÇOCUK İSTİSMARI DAVALARININ SONUÇLARINA ETKİ EDEN FAKTÖRLERİN BELİRLENMESİ İÇİN MAKİNE ÖĞRENMESİ YAKLAŞIMLARI

**DOKTORA TEZİ**

**Saime Şule AKSAKAL**

**Enstitü Anabilim Dalı : İstatistik**

**Bu tez 14/05/2025 tarihinde aşağıdaki jüri tarafından oybirliği ile kabul edilmiştir.**

| | | |
|---|---|---|
| **Prof. Dr.**<br>**Haşim ÇAYIR**<br>**Jüri Başkanı** | **Prof. Dr.**<br>**Erol EĞRİOĞLU**<br>**Üye** | **Prof. Dr.**<br>**Eren BAŞ**<br>**Üye** |
| **Doç. Dr.**<br>**Nihat TAK**<br>**Üye** | | **Dr. Öğr. Üyesi**<br>**Erdinç YÜCESOY**<br>**Üye** |

**Prof. Dr.**
**Bahadır KOZ**
**Enstitü Müdürü**

# BEYAN

Tez içindeki tüm verilerin akademik kurallar çerçevesinde tarafımdan elde edildiğini, görsel ve yazılı tüm bilgi ve sonuçların akademik ve etik kurallara uygun şekilde sunulduğunu, kullanılan verilerde herhangi bir tahrifat yapılmadığını, başkalarının eserlerinden yararlanılması durumunda bilimsel normlara uygun olarak atıfta bulunulduğunu, tezde yer alan verilerin bu üniversite veya başka bir üniversitede herhangi bir tez çalışmasında kullanılmadığını beyan ederim.

Saime Şule AKSAKAL

14/05/2025

# TEŞEKKÜR

Bu çalışmanın her safhasında bana yol gösteren ve yardımlarını esirgemeyerek bana karşılaştığım her zorluğun üstesinden gelmemde yardımcı olan çok değerli tez danışman Hocam Sayın Prof. Dr. Erol EĞRİOĞLU' a en içten saygı ve teşekkürlerimi sunarım.

Çalışmalarım sırasında bilgi ve tecrübelerinden yararlandığım, desteğini ve yardımını esirgemeyen Sayın Prof. Dr. Eren BAŞ'a, Sayın Dr. Öğr. Üyesi Erdinç YÜCESOY'a ve çalışmamın zorlu safhalarında hem dostluğunu hem bilgilerini benden esirgemeyen kıymetli hocam Prof. Dr. Haşim ÇAYIR'a en içten saygılarımla teşekkür ederim.

Ayrıca Giresun Üniversitesi İstatistik Bölümü'nün çok değerli akademisyenlerine sonsuz saygı ve sevgilerimi iletirim.

Akademik çalışmalarıma yeniden dönmem konusunda bana ilham veren ve bu kadar hassas bir konuya katkı sunmamda büyük etkisi bulunan UCİM Derneği Başkanı Sayın Saadet ÖZKAN'a, Başkan Yard. Sayın Yücel CEYLAN'a ve Türkiye'nin her ilinde çocuklar için mücadele eden kıymetli UCİM gönüllülerine yürekten saygı ve sevgilerimle teşekkür ederim.

Şüphesiz bugünlere gelmemde en büyük emeği olan, en zorlu zamanlarda dualarını ve her türlü desteklerini benden esirgemeyen anneme ve babama en içten dileklerimle teşekkür ederim.

Varlığı, anlayışı ve bana verdiği güç için kızıma sonsuz teşekkürler ederim.

# İÇİNDEKİLER

# SİMGELER VE KISALTMALAR LİSTESİ

| | |
|---|---|
| UNICEF | : Birleşmiş Millletler Uluslararası Çocuklara Acil Yardım Fonu |
| UNFPA | : Birleşmiş Millletler Uluslararası Nüfus Fonu |
| ASHB | : Aile ve Sosyal Hizmetler Bakanlığı |
| UCIM | : Saadet Öğretmen Çocuk İstismarı ile Mücadele Derneği |
| CHAID | : Ki-Kare Otomatik Etkileşim Belirleyicisi |
| AID | : Otomatik Etkileşim Belirleyicisi |
| DVM-SVM | : Destek Vektör Makineleri (Support Vector Machines) |
| mRMR | : Minimum Fazlalık, Maksimum İlgililik |
| ANOVA | : Varans Analizi |
| kNN | : K-En Yakın Komşuluk |
| CART | : Sınıflandırma ve Regresyon Ağaçları |
| CHI2 | : Ki-Kare Bağımsızlık veya Uyum Testi |
| YSA | : Yapay Sinir Ağları |
| KA | : Karar Ağacı |
| RO(RF) | : Rastgele Orman (Random Forest) |
| NB | : Naive Bayes Algoritması |
| LR | : Lojistik Regresyon |
| DÖ | : Derin Öğrenme |
| TBA | : Temel Bileşenler Analizi |
| TPR | : Doğru Pozitif Oranı (Duyarlılık) |
| FNR | : Yanlış Negatif Oranı |
| PPV | : Pozitif Tahmin Değeri (Kesinlik) |
| FDR | : Yanlış Keşif Oranı |
| ROC | : Alıcı İşletim Karakteristik Eğrisi |
| AUC | : Eğri Altında Kalan Alan |

# ŞEKİLLER LİSTESİ

# TABLO LİSTESİ

# ÇOCUK İSTİSMARI DAVALARININ SONUÇLARINA ETKİ EDEN FAKTÖRLERİN BELİRLENMESİ İÇİN MAKİNE ÖĞRENMESİ YAKLAŞIMLARI

## ÖZET

Modern bilgi teknolojisi, bilimsel ve sosyal araştırma verilerinin toplanmasını ve saklanmasını mümkün kılmaktadır. Bazı istatistiksel yöntemler, veriler arasındaki mevcut veya gizli ilişkilerin ortaya çıkarılmasında gerekli varsayımlar sağlandığında oldukça güvenilir sonuçlar verebilmektedir. Ancak gerçek hayattan toplanan veriler çoğu zaman bu varsayımları karşılamadığından, tahmin için daha az varsayım gerektiren, esnek ve karmaşık veri setlerine uygulanabilen veri madenciliği yöntemleri geliştirilmiştir. Verileri işlemek ve anlamlı bilgiler üretmek için veri madenciliği tekniklerini içeren makine öğrenmesi yöntemlerinin kullanımı son yıllarda yaygınlaşmıştır. Bu tez çalışmasının amacı çocuk istismarı davalarının sonucuna etki edebilecek özniteliklerin belirlenmesinin yanı sıra, makine öğrenmesi algoritmalarının performanslarını karşılaştırarak, en iyi modelin özelliklerinin tahmin edilmesidir. Çalışmanın örneklemini UCİM Saadet Öğretmen Çocuk İstismarı ile Mücadele Derneği'nin katılım talebinde bulunduğu 61 çocuk istismarı vakasının verileri oluşturmaktadır. Çalışmanın bağımlı değişkeni, sanığın yargılama sonunda ceza alıp almaması olarak belirlenmiştir. Uygulamada modellere ve değişkenlerin önem değerlendirmelerine dayalı doğruluk, precision(kesinlik), recall(duyarlılık), F1 skoru, ROC eğrisi ve AUC değeri gibi farklı metrikler dikkate alınmıştır. Çalışmada karar ağaçlarının bir uygulaması olan CHAID algoritması, destek vektör makineleri, lojistik regresyon analizi, rastgele orman, çoklu algoritmalarla full model ve öznitelik seçimi modellerinin performansları karşılaştırılmıştır. Bu çalışmada uygulanan tekniklerden CHAID, lojistik regresyon, DVM (Radial, Fine Gaussian, Quadratic), RO ve çoklu algoritma ile elde edilen performanslar oldukça yüksek ve birbirlerine çok yakın değerler olsa da en yüksek performansı gösteren yöntemin rastgele orman algoritması olduğu ayrıca öznitelik seçiminin modelin performansını arttırdığı tespit edilmiştir. Dava sonuçlarına en çok etki eden faktörlerin "UCİM' in davaya müdahilliği, ihbarcının çocukla yakınlığı, dava sırasında sanığın tutukluluk durumu ve istismarın gerçekleştiği yer" olduğu ortaya çıkmıştır. Sonuç olarak, bu çalışma hem makine öğrenimi uygulamaları konusunda uzman sistemler üzerinde çalışan uygulayıcılar ve araştırmacılar için hem de çocuklarla ilgili hukuk, tıp, psikoloji, çocuk gelişimi ve eğitimi üzerine çalışan uzmanlar için değerli olacaktır.

**Anahtar kelimeler:** Makine Öğrenmesi, Öznitelik Seçimi, Lojistik Regresyon, Destek Vektör Makineleri, CHAID Algoritması, Rastgele Orman, Çocuk İstismarı Davaları.

# MACHINE LEARNING APPROACHES FOR DETERMINING FACTORS AFFECTING THE OUTCOMES OF CHILD ABUSE CASES


## SUMMARY

Modern information technology enables the collection and storage of scientific and social research data. Some statistical methods can provide highly reliable results in uncovering existing or hidden relationships between data when the necessary assumptions are met. However, since real-world data often do not satisfy these assumptions, data mining methods that require fewer assumptions and can be applied to flexible and complex datasets have been developed for prediction purposes. The use of machine learning methods, which incorporate data mining techniques to process data and generate meaningful insights, has become widespread in recent years. The aim of this thesis is not only to identify the attributes that may influence the outcomes of child abuse cases but also to compare the performance of machine learning algorithms and predict the characteristics of the best-performing model. The sample of this study consists of data from 61 child abuse cases in which the UCİM Saadet Öğretmen Association Struggling with Child Abuse requested involvement. The dependent variable of the study is whether the defendant received a sentence at the end of the trial. For this comparison, various metrics such as accuracy, precision, recall, F1 score, ROC curve, and AUC value, based on algorithm models and variable importance assessments, have been considered. The study compares the performances of the CHAID algorithm (a decision tree implementation), support vector machines, logistic regression analysis, random forest, multiple algorithms with full models, and feature selection models. Among the techniques applied in this study, although the performances of CHAID, logistic regression, SVM (Radial, Fine Gaussian, Quadratic), RF, and multiple algorithms were all quite high and close to each other, the random forest algorithm was found to be the best-performing method. Additionally, it was determined that feature selection improved the model's performance. The factors that most significantly influenced case outcomes were identified as "UCİM's involvement in the case, the informant's relationship with the child, the defendant's detention status during the trial, and the location where the abuse occurred". In conclusion, this study will be valuable not only for practitioners and researchers working on expert systems in machine learning applications but also for professionals in law, medicine, psychology, child development, and education related to children.



**Keyword:** Machine Learning, Feature Selection, Logistic Regression, Support Vector Machines, CHAID Algorithm, Random Forest, Child Abuse Cases.

# BÖLÜM 1.GİRİŞ

Çocuk cinsel istismarı, Dünya Sağlık Örgütü tarafından bir çocuğun bir ebeveyn, bakım veren ya da yetişkin tarafından cinsel uyarılma veya haz için kullanılması olarak tanımlanmaktadır [1]. Bu durum, çocukta fizyolojik, psikolojik ve duygusal olarak derin travmalara yol açmaktadır. Çocuklar cinsel saldırı, teşhircilik, çocuk pornografisi ve çocuk ticareti yoluyla mağdur edilebilmektedir. UNICEF' in Ekim 2024' te yayınladığı rapora göre çocuk yaşta fiziksel temas içeren cinsel istismara kalmış kız çocuğu sayısı yaklaşık 340 milyon, erkek çocuk sayısı ise 240 ile 310 milyon aralığında olup bu sayı her 8 kız çocuktan birinin ve her 11 erkek çocuğundan birini işaret etmektedir. Temas içermeyen cinsel istismar da dahil edildiğinde bu sayının kız çocuklarında 650 milyona, erkek çocuklarında ise 410 ile 530 milyon arasına yükselmekte olduğu ve dünyadaki her 5 kişiden birinin çocukluk çağında cinsel istismara maruz kaldığı belirtilmiştir [2]. Birleşmiş Milletler Nüfus Fonu (UNFPA) ve Hacettepe Üniversitesi'nin 2020 yılında iş birliğiyle yayımladığı rapora göre, dünya genelinde her 5 çocuktan biri evlendirilmiştir ve Türkiye'de 18-45 yaş arasındaki kadınlar için de aynı oran geçerlidir [3].

Çocukların gelişimsel dönemlerine bağlı olarak istismar eyleminin farkında olmaması, farkında olan çocukların bazen yaşadıkları istismar için kendilerini suçlamaları, istismarcı tarafından tehdit edilmeleri ya da ailelerinin tepkisinden korkmaları nedeniyle, çocuklara yönelik cinsel istismar vakalarının yalnızca onda birinin resmi makamlara yansıdığı tahmin edilmektedir [4]. Resmi istatistikler kız çocuklarının daha sık cinsel istismara maruz kaldığını gösterse de erkek çocukların cinsel istismar vakalarını sosyokültürel yaptırımlar nedeniyle açıklayamaması, doğru rakamlara ulaşmayı zorlaştırmaktadır. Genel kanının aksine, istismarcılar belirli bir profile sahip değildir ve çeşitli sosyodemografik, ekonomik ve kültürel düzeylerden gelebilirler. Yine yaygın inanışın aksine, istismarcılar genellikle çocuğun tanıdığı kişilerdir. Bu

durum, çocuğun yaşadığı travmayı ağırlaştırabilir ve üstesinden gelmesini zorlaştırabilir.

Türkiye'de çocukların korunmasına yönelik yasalar ve düzenlemeler arasında Türk Medeni Kanunu, 6284 sayılı Ailenin Korunması ve Kadına Karşı Şiddetin Önlenmesine Dair Kanun, Türk Ceza Kanunu ve Çocuk Koruma Kanunu bulunmaktadır.

6284 sayılı Kanun, çocukların ve kadınların aile içi şiddet ve cinsel istismardan korunmasında çok önemli bir yere sahiptir. Türk Ceza Kanunu'nun 5237 sayılı yasasında çocuk cinsel istismarına yönelik ceza hükümleri yer almaktadır. 5395 sayılı Çocuk Koruma Kanunu'nun amacı, korunmaya muhtaç ya da suça sürüklenmiş çocukların korunmasına ilişkin usul ve esasları düzenlemek, çocukların haklarını ve refahını güvence altına almaktır [5].

Türkiye'de çocuk ihmali ve istismarı üzerine çalışan kamu kurum ve kuruluşları arasında kolluk kuvvetlerinin yanı sıra Çocuk Koruma Kurumları, Çocuk Destek Merkezleri, Çocuk İzlem Merkezleri ve sivil toplum kuruluşları bulunmaktadır. Ebeveynlerini kaybeden, bakım verenleri bulunmayan, aileleri tarafından yeterli bakım sağlanmayan ya da aileleri tarafından istismara uğrayan çocukların bakımı ve gözetimi, Aile ve Sosyal Hizmetler Bakanlığı'na bağlı Çocuk Koruma Kurumlarında sağlanmaktadır. Suça maruz kalan, suça sürüklenen, sokakta yaşamak zorunda kalan ve risk altındaki çocuklara ise Çocuk Destek Merkezlerinde sosyal hizmetler sunulmaktadır.

Cinsel istismara maruz kaldığından şüphelenilen çocukların farklı adli mercilere tekrar tekrar ifade vermek zorunda kalarak yaşadığı travmanın önlenmesi ve adli muayene ile diğer tüm işlemlerin tek bir merkezde gerçekleştirilmesi amacıyla, Sağlık Bakanlığı'na bağlı Çocuk İzlem Merkezleri kurulmuştur. Sokakta yaşamak zorunda kalan veya yasa dışı durumdaki terk edilmiş çocukların tespit edilmesi ve ilgili kurumlara yönlendirilmesi amacıyla, ülke genelindeki 81 ilde Çocuk Polisi birimleri faaliyet göstermektedir [6].

Çocuk ihmal ve istismarı konusunda çalışmalar yapan diğer bir alan ise sivil toplum kuruluşlarıdır. Sivil toplum, bireylerin gönüllü olarak bir araya gelerek toplumsal bir sorunu çözmek için çaba gösterdiği oluşumlardır. Türkiye'de devlet politikalarını destekleme misyonuyla çocuk ihmali ve istismarı ile mücadele eden sivil toplum kuruluşlarından biri de UCIM (Saadet Öğretmen Çocuk İstismarı ile Mücadele Derneği)'dir. Derneğin kurucu başkanı Saadet Özkan, İzmir'de çalıştığı bir köy ilkokulunda okul müdürünün öğrencileri istismar ettiğini fark etmiş ve durumu ilgili makamlara bildirmiştir. Saadet Öğretmen'in kamu desteğiyle çocukların yanında yer aldığı bu yargı süreci sonucunda, sanık 82 yıl 6 ay hapis cezasına çarptırılmıştır. Bu olayın ardından Saadet Özkan, iş insanı Yücel Ceylan ile 2017 yılında UCIM derneğini kurarak çocuk ihmali ve istismarıyla mücadeleye başlamıştır.

UCIM, 8 yıldır Türkiye'nin 55 ilinde gönüllüleriyle aktif olarak çalışmaktadır ve 2020 yılında İzmir'de Avrupa'nın ilk Çocuk İhmal ve İstismar Önleme Ofisi'ni kurmuştur. 2025 yılı itibarıyla bu ofislerin sayısını 14'e çıkarmıştır. Çocuk istismarını önlemenin en önemli yolunun eğitim olduğu düşüncesiyle çocuklar, ebeveynler, yetişkinler, öğretmenler, kolluk kuvvetleri ve çocukla temasta bulunabilecek tüm kişi ve kurumlara yönelik çocuk hakları odaklı eğitimler düzenlemektedir. Ankara, Antalya, Denizli, Diyarbakır, Erzurum, Giresun, İstanbul, İzmir, Konya, Mersin, Niğde, Van, Tekirdağ ve Trabzon'da faaliyet gösteren önleme merkezlerinde psikologlar, sosyologlar, çocuk gelişimi uzmanları, psikolojik danışmanlar ve sosyal hizmet uzmanları görev yapmaktadır. Dernek, çocuk hakları farkındalığını il ve ilçe merkezlerinden köylere kadar taşımaktadır.

Ayrıca UCIM' in rehberlik ve yönlendirme ihbar hattında görevli çocuk ruh sağlığı profesyonelleri, mağdur çocuklara ve ailelerine rehabilitasyon ve sağaltım konusunda 7 gün 24 saat gönüllü olarak destek vermektedir [7]. UCIM hukuk koordinatörlüğü ihbar hattı ise her an hizmet veren nöbetçi avukatların bulunduğu bir destek sistemidir. İhbar geldiğinde gönüllü avukatlar öncelikle mağdur çocuk ve ailesini ilgili devlet kurumlarına yönlendirerek ihbarı resmileştirmekte, sonrasında ise tüm adli süreç boyunca davayı takip ederek mağdurları desteklemektedir.

UCIM' in takip ettiği ya da müdahil olduğu çok sayıda çocuğun cinsel istismarı davasında mağdur çocuğu ve ailesini desteklemiş olan UCIM hukuk müşavirliğinin görüşleri ile, dava sonuçlarına etki edebilecek faktörler şu şekilde belirtmiştir;

- **Çocuğun Yaşı** Özellikle mağdur çocuğun yaşının istismarı fark edemeyecek kadar küçük olduğu davalarda çocuğun kendini ifadede yetersiz kalması nedeniyle dava süreci önemli ölçüde etkilenebilmektedir.

- **İstismarın Sürekliliği** Sistematik saldırıların genel olarak çocuğun ailesinden veya yakın çevresinden biri tarafından gerçekleştirilmesi, çocuğun güvenli alanından saldırıya uğraması gerekçesiyle dava sonucunu önemli ölçüde etkileyebilmektedir.

- **İstismarın Gerçekleştiği Yer** Eylemin ev ortamında gerçekleşmesi, çocuğun yardım isteyememesi ya da duyulmaması, iddiaların temelsiz olduğu yönünde bir algı yaratabilmektedir.

- **Bildirimin Zamanlaması** İstismar bildirimlerinin hemen yapılması, ifadelerin geçerliliğini koruması ve detayların doğru aktarılabilmesi açısından önemlidir. Özellikle fiziki delillerin kaybolmaması için ilk 72 saat içinde bildirimde bulunulmalıdır. Gecikmiş bildirimlerde çocuklar bazı detayları unutabilmekte veya kendi kendilerini yatıştırmak için hayali detaylar üretebilmektedir.

- **Failin Çocukla Yakınlık Derecesi** Failin aile bireyi olması durumunda çocuk, aile bütünlüğünün bozulmasından veya failden korkarak ifadesini değiştirebilir. Bu, failin daha az ceza almasına neden olabilir.

- **İhbarcı ile Çocuğun Yakınlık Derecesi** İstismarı ihbar eden kişinin aileyle bir bağının olmaması, tarafsız ve objektif bir ihbar algısı yaratabilir.

- **İstismar Türü** Nitelikli cinsel istismarın kanıtlanabilmesi için ciddi adli deliller gerekmektedir. Niteliksiz cinsel istismarda ise genellikle fiziksel bulgu olmadığı için soruşturma aşamasının ciddiyetle sürdürülmesi gerekmektedir.

- **Soruşturma Aşaması** Dava sürecinin sağlıklı işlemesi, soruşturma aşamasında belgelerin, bulguların ve ipuçlarının eksiksiz toplanmasına bağlıdır.

- **Mağdur Çocuğun İlk İfadesinin Uzman Gözetiminde Alınması** Travmaya maruz kalan çocuğun ilk ifadesinin bir çocuk psikolojisi profesyoneli gözetiminde alınması, daha sağlıklı bilgi elde edilmesini ve sürecin doğru ilerlemesini sağlar.

- **Mağdur Çocuğun Dava Sürecinde Faille Karşılaşması** Duruşma sırasında mağdurun faille karşılaşması, çocuğun korkmasına, ifadesini değiştirmesine ve tekrar travmatize olmasına neden olabilir. 2016 yılında 9 yaşındaki Y., duruşma öncesi sanıkla duruşma salonunda karşı karşıya geleceği ve yüzleşeceği kaygısı ile kalp krizi geçirerek duruşmadan 2 gün önce yaşamını yitirmiştir.

- **Tutukluluk, Adli Kontrol ve Uzaklaştırma Tedbirlerinin Alınması** Dava süreci boyunca mağdur çocuğun ve ailesinin korunması için adli kontrol, tutuklama ve uzaklaştırma tedbirlerinin uygulanması önemlidir. Aksi takdirde fail, mağduru, ailesini, ihbar eden kişiyi veya tanıkları tehdit ederek veya şiddet göstererek davanın seyrini değiştirebilmektedir.

- **Davaya Kamuoyu, Aile Sosyal Hizmet Müdürlükleri ve UCIM' in Desteği** Kamuoyu desteği ve farkındalık, dava sürecini etkileyen önemli unsurlardır. Kamu vicdanının yanı sıra Aile ve Sosyal Hizmet Bakanlığı ile UCIM' in hukuk profesyonellerinin davalara katılımı da bu anlamda önemli bir rol oynamaktadır.

# BÖLÜM 2. LİTERATÜR ÖZETİ

Çocuk istismarı davaları ile ilgili alınan uzman görüşleri doğrultusunda bilimsel kaynaklar araştırılmıştır. Çocuk ihmal ve istismarı ile ilgili literatürde çeşitli disiplinlerden çok sayıda çalışma yapılmış olsa da çocuk istismarı davalarının sonuçlarına etki edebilecek faktörleri inceleyen herhangi bir çalışma bulunmamaktadır. Civelek' in [8] yaptığı tez çalışmasına göre 353 tez çalışmasından %51,6'sı istismar tanımı, %35,1'i sonuçlarını inceleme, %7,1'i müdahale, %5,4'ü önleme, %0,8'i tedavi-rehabilitasyon konuları hakkında yazılmıştır. Bu çalışmalar arasında Pınar'ın yaptığı tez çalışması [9] faktör analizi yöntemiyle incelenen bir ihmal ve istismar farkındalık çalışmasıdır.

Çocuk ihmal ve istismarı konusu toplumsal ve vicdani önemi nedeniyle bilim dünyasında da son zamanlarda üzerinde sıklıkla durulan bir konu olmuştur. Yüce ve ark. [10], 2010-2018 yılları arasında Nevşehir' de sağlık kurumlarına adli yollarla getirilen 202 çocuk cinsel istismarını vakalarını incelemiş, risk faktörlerini belirlemiştir. Uslu [11], 2012-2017 yılları arasında gerçekleşmiş ve karara bağlanmış 142 çocuk istismarı davasını incelemiş ve çocuk istismarındaki risk faktörlerini sosyo-demografik açıdan analiz etmiştir. Aydın ve ark. [12] Ocak 2006- Aralık 2012 tarihleri arasında On dokuz Mayıs Üniversitesi Tıp Fakültesi Adli Tıp Anabilim Dalı'na sevk edilen 1002 çocuk cinsel istismarı vakasını incelemiş ve vakalar üzerine tanımlayıcı analizler yaparak risk faktörlerini belirlemiştir. Çubuk ve Şeker [13], yaptıkları literatür incelemesi ile çocuk ihmal ve istismarına yönelik çalışmaların büyük çoğunluğunun bilgi ve farkındalık ölçme ya da risk faktörlerini belirleme üzerine olduğunu belirtmişlerdir.

Tüm bu çalışmalar ışığında çocuk ihmal ve istismarı konusunda tanımlayıcı, risk belirleyen ya da farkındalık ölçen çalışmaların yanı sıra çocuk cinsel istismarı davalarının süreç ve sonucunu etkileyebilecek faktörlere, mağdur edilen çocukların

dava sırasında uğradıkları hak ihlallerine, önleme çalışmaları kapsamında çocuk hakları konusunun yaygınlaştırılmasına, istismar sonrası sağaltım ve rehabilitasyon süreçlerine ışık tutacak bilimsel çalışmaların yapılması gerektiği görülmektedir.

Veri madenciliği, veri erişiminin toplanmasının ve depolanmasının çok kolay olduğu teknoloji çağında, büyük veri yığınları içinden bilgiye ulaştıracak saklı veya açık verilerin türlü analiz teknikleri ile ortaya çıkarılması sürecidir. Yapay zekâ kavramı insanlar gibi öğrenen, genelleyebilen, bağlantılar kurabilen ve ilerde insan zekasından bağımsız olarak gelişebileceği düşünülen makinelerdir. Bilimsel anlamda makine öğrenmesi kavramı ise yapay zekâ arayışı esnasında karşımıza çıkmaktadır. Makine öğrenmesi, genel olarak veri madenciliği tekniklerini kullanarak veriler üzerinden elde edilen bilgilerle temellendirdiği algoritmaları, açıkça görevlendirmeye gerek kalmadan bilinmeyen veriler üzerine geliştiren bir sistemdir.

Makine öğrenmesinin uygulandığı bilim dalları oldukça geniş bir yelpazeye yayılmıştır. Literatür tarandığında makine öğrenmesinin uygulandığı mühendislik, endüstri, işletme, ekonomi, borsa uygulamaları, sağlık, bilim teknolojileri, müzik, mimarlık, eğitim, felsefe gibi bilim dallarından çok sayıda çalışma yapıldığı görülmüştür. Uygulandığı konulara örnek olarak doğal dil işleme, arama motorları, bilgisayarlı görmede nesne tanıma, biyoinformatik, DNA dizilerinin sınıflandırılması, borsa çözümlemesi, tıbbi tanı, kredi kartı dolandırıcılığı denetimi, makine algılaması, beyin-makine arayüzleri, robot gezisi, konuşma ve el yazısı tanıma, sözdizimsel örüntü tanıma, oyun oynama ve uyarlamalı web siteleri verilebilir.

Tez çalışmasında çocuk cinsel istismarı davalarında sanıkların ceza alıp almadıklarına dair sınıflama tahmini yapıldığı için makine öğrenmesi teknikleri arasından ikili sınıflandırma yöntemleri tercih edilmiş ve bu yönde açıklamalar yapılmıştır. Bu yöntemler ağaç tabanlı yöntemler (CHAID algoritması, rastgele orman algoritması), lojistik regresyon, destek vektör makineleri (quadratik DVM ve fine Gaussian DVM), öznitelik seçimi algoritmaları (mRMR, ki-kare, ReliefF, ANOVA, Kruskal-Wallis) olup her yöntemin çalışma şekline dair açıklamalar yapılarak yöntemler doğruluk ve tahmin performansları açısından kıyaslanmıştır.

Makine öğrenmesi kavramını Arthur Samuel 1959 yılında kullanmıştır [14]. Sınıflandırma problemlerinde ise gelişmiş özellikleri nedeniyle makine öğrenmesi son derece popüler bir alandır. Karar ağaçları, sınıflandırma problemlerinde başvurulan bir denetimli öğrenme yöntemidir. Karar ağacı algoritmalarından CHAID, Kass [15] tarafından geliştirilmiştir. Mümkün olan en az değişken sayısı ile bağımlı ve bağımsız değişken arasındaki ilişkiyi tanımlayan modeli oluşturmak üzerine kurgulanan lojistik regresyon analizi, ilk defa 20. yüzyılın ilk yarısında Berkson tarafından önerilip farklı alanlarda kullanımı sağlanmıştır [16]. Vladimir N. Vapnik'in 1995 yılında geliştirdiği destek vektör makineleri sınıflandırma ve eğri uydurma araştırmaları için oldukça başarılı sonuçlar veren bir makine öğrenmesi tekniğidir [17]. Araştırma modeline göre eğitim aşaması esnasında karar ağaçları oluşturarak regresyon veya sınıflama tahmini yapan rastgele orman algoritması Ho [18] tarafından önerilmiş, Breiman [19] tarafından geliştirilmiştir. Makine öğrenmesinde model performansını artırmak için gerekli veri ve özelliklerin seçimi üzerine geliştirilmiş öznitelik seçimi (feature selection), temellerini Fisher Lineer Diskriminant Analizi yönteminden almıştır [20]. Makine öğrenmesi teknikleri anlaşılabilirliği ve uygulamada sağladığı kolaylıklar nedeniyle günümüzde sınıflama ve regresyon modelleri için sıklıkla tercih edilmektedir. Şata ve Çakan [21], lise öğrencilerinin fizik dersine yönelik tutumlarını etkileyen değişkenleri belirlemek için CHAID analizi ve lojistik regresyon analizi tekniklerini karşılaştırmıştır. Baraklı ve Küçüker [22], karar destek makineleri ve rastgele orman yöntemleri ile vücut yağ yüzdesi tahmini yapmıştır. Izgara tarama teknikleri ile belirlenmiş yeni regresyon parametreleri ile tahmin doğruluk değerleri artırılmıştır. Uçkan ve Karabulut [23] çıkarımsal metin özetlemede makine öğrenmesi yöntemlerinin uygulanabilirliğini, etkinliğini test etmiş ve performanslarını karşılaştırmıştır. Doğan ve Özdamar [24], 400 ailenin aile planlamalarına etki edebilecek faktörleri CHAID algoritması üzerinden incelemişlerdir. Güvenç ve ark. [25], bilişim dersleri üzerinden öğrencilerin dönem sonu başarılarını tahmin ederek başarısızlığa etki edebilecek faktörleri çeşitli makine öğrenme teknikleri ile tespit etmiştir. Gür ve Tarhan Mengi [26], bankacılıkta hile tespiti için birçok makine öğrenmesi yöntemini karşılaştırmış ve en başarılı sonuca karar ağaçları yöntemi ile ulaşmışlardır. Alkan [27], öğrencilerin sınav performanslarını tahmin etmek için

denediği teknikler arasından doğru sınıflama oranı en yüksek tekniğin destek vektör makineleri olduğunu tespit etmiştir. Ercan [28], destek vektör regresyonu ve yapay sinir ağları yöntemlerini karşılaştırarak konut kira fiyat tahmini yapmıştır. Serdarer Kuzu ve Giray Yakut [29], çalışmalarında finansal başarısızlık tahmini yapmak için destek vektör makinelerinden yararlanmışlardır. Şahin ve ark. [30], hamilelik sürecinde anne sağlığına etki edebilecek risk faktörlerini belirlemek için makine öğrenmesi yöntemlerinden yararlanarak model performanslarını karşılaştırmışlardır. Korkmaz ve ark. [31], yaptıkları çalışmada Botnet Tespiti için sınıflandırma ve regresyon ağaçları ile rastgele orman algoritmasını kullanmış, öznitelik seçimi (feature selection) ile model performansını artırmışlardır. Kayalı ve Savaş [32], öğrenci performansını etkileyen temel faktörleri incelemiş, bu faktörlerden en yüksek etkiye sahip olanların ebeveynlerin eğitim düzeyleri, öğrencinin daha önce aldığı eğitim kalitesi ve ailenin gelir düzeyi olduğunu tespit etmişlerdir. Kasım [33] yüksek lisans tezinde karar ağaçları, rastgele orman, lojistik regresyon, k-en yakın komşu (kNN), Naive Bayes ve yapay sinir ağları algoritmalarını kullanılarak müşteri kayıp analizi yapmıştır. Budak [34], araştırmasında kar marjı öngörüsü için CHAID, CART ve doğrusal regresyon yöntemlerini kullanılmış, modeller arasında en yüksek doğru sınıflama tahminini CHAID yöntemiyle elde etmiştir. Durnagöl [35], çevrimiçi yayın yapan bir medya kuruluşunun haber ve raporlarının sınıflandırılması üzerine çalışmıştır. Kullandığı makine öğrenme teknikleri arasında en yüksek performansı rastgele orman yöntemi göstermiştir. Selvi [36], liseye başlayacak olan öğrencilerin başarılarına etki edebilecek demografik özellikleri öznitelik seçimi ve rastgele orman teknikleri ile araştırmıştır. Çam [37], tez çalışmasında firmaların verilerini daha etkili kullanması yoluyla karar verme aşamasını daha etkin hale getirmek üzere bir sistem geliştirmiştir. Aslan [38], çalışmasında çeşitli Sıralı Küme Örneklemesi algoritmalarının model performanslarını karşılaştırmıştır. Karaatlı [39], çalışmasında asansörlerin arıza nedenlerini sınıflandırmak için çeşitli teknikler denemiş, Iterative Classifier Optimizer ve Logitboost algoritmalarının en yüksek performansa sahip olduğunu bulgulamıştır. Yılmaz [40], tez çalışmasında Destekleme ve Yetiştirme Kursu programında birlikte öğrenme tekniği kullanımının öğrencilerin akademik başarısına etkisini incelemiştir. Köksal [41], İstanbul Marmara Üniversitesi Araştırma Hastanesi'ne gelen COVID-19 vakaları üzerinde oksijen ihtiyacı ve yoğun bakım

ihtiyacı olmak üzere belirlenen iki farklı sonlanım durumu ile ilk laboratuvar sonuçları arasındaki ilişkiyi k-En yakın komşu, torbalama, rastgele orman ve karar ağacı makine öğrenmesi yöntemleri kullanılarak analiz etmiştir. Sevgen [42], hazırladığı yüksek lisans tezi ile k en yakın komşuluk, yapay sinir ağlar, destek vektör regresyonu, rasgele orman, çoklu regresyon analizi algoritmalarını kullanarak 4 farklı gayrimenkul türünün değerliliğini etkileyen faktörleri belirleyip model performanslarını karşılaştırmıştır. Engin ve İlter Fakhouri [43], yaptıkları araştırma ile işletmelerin nakit akış tablolarını kullanarak finansal risklerini tahmin etmeyi amaçlamaktadır. Bu amaçla, Aşırı gradyan artırma (XGBoost), gradyan yükseltme (Gradient Boosting) ve yapay sinir ağları (Neural Network) gibi makine öğrenmesi algoritmaları karşılaştırılmıştır. Çalışmanın sonuçlarına göre, XGBoost algoritması %80 doğruluk skoru ile en başarılı model olarak belirlenmiştir. Serdarer Kuzu [44], "Makine öğrenmesi algoritmaları ile LGS başarısı tahmin modelleri kurulması" başlıklı çalışmasında, öğrencilerin liselere geçiş sistemi başarısını öngörmek için Naive Bayes, destek vektör makineleri, karar ağacı ve lojistik regresyon gibi sınıflandırma algoritmaları ile tahmin modelleri oluşturarak model performanslarını karşılaştırmıştır. Bilenler [45], tezinde Kaggle platformu üzerinden alınan float türü verileri karar ağaçları tekniklerinden CART algoritması ile sınıflandırma çalışması yapmıştır. Mısırlıoğlu ve ark. [46], Kanser Genomu Atlas Kolon Adenokarsinom Koleksiyonu veri kümesinden klinik ve genomik veriler kullanan CRC hastalarında sağkalım tahmini için yapay zekâ destekli bir klinik karar destek sistemi geliştirme amacıyla karar ağacı, destek vektör makineleri (DVM), rastgele orman ve Naive Bayes yöntemleri ile kurulan modelleri doğruluk ve performansa göre değerlendirmiştir. Karagöz [47], çalışmasında seçilen paylar için kapanış fiyatına etkisi yüksek olan finansal değişkenler ve makine öğrenimi algoritmalarını araştırmıştır. Korkmaz [48], doktora tezinde İstanbul'da gerçekleşen trafik kazalarını makine öğrenmesi algoritmaları ile değerlendirip kaza sürelerinin tahmini için çeşitli algoritmaların performanslarını karşılaştırmıştır. Alan [49], makine öğrenmesi algoritmalarının matematiksel temellerini ve bu algoritmaların farklı veri setleri üzerindeki performanslarını incelemiştir. Çalışmada, k-en yakın komşu (kNN), Naive Bayes, karar ağaçları, lojistik regresyon ve destek vektör makineleri (DVM) gibi sınıflandırma algoritmaları ele alınmıştır. Bu algoritmaların performansları, UCI

Machine Learning Repository'den alınan ve boyut, büyüklük, veri tipi açısından farklılık gösteren üç farklı veri seti (Mushroom, Congressional Voting Records, Tic-Tac-Toe) üzerinde değerlendirilmiştir. Karakullukçu [50], çalışmasıyla yanık hastalarına ilişkin 105 adet dijital (2D) görüntüdeki sağlıklı ve yanık deriye ilişkin bölgelerin tespitini araştırmıştır. Günerkan [51], gümrük sistemlerinde insan kaynaklı hataların tespitini sağlayarak hata oranlarının en düşük seviyeye indirgenmesi amacıyla makine öğrenmesi yöntemlerinden yararlanmıştır. Maden [52], karar ağaçları, k -en yakın komşuluk, Naive Bayes, yapay sinir ağları, rastgele orman tekniklerini karşılaştırarak tıp fakültesi kurul sınavları sonuçlarını tahmin edebilecek bir model geliştirmiştir. Işık [53], biyosentez ve metabolik süreçleri henüz tam olarak açıklanmamış, insan metabolizmasında önemli bir aminoasit/protein olan sitrülin için Bayes sınıflandırma algoritması üzerinden R Yazılımı ile bir arayüz geliştirmiştir. Vupa Çilengiroğlu ve Genç [54], çalışmalarında bir firmanın toplam ekipman etkinlik puanına etki edebilecek değişkenleri karar ağaçları, lojistik regresyon ve yapay sinir ağları yöntemleri ile karşılaştırmışlardır. Doğan [55], tez çalışmasında finansal başarısızlık tahmini için destek vektör makineleri yöntemi ile yeni bir model araştırarak tahmin başarısını geliştirmeye çalışmıştır. Aksoy [56], çalışmasında İstanbul Borsasındaki 88 şirketin 2000-2019 yılları arasında finansal işlemlerinde hile olup olmadığını bir yıl önceden tahmin edebilmek için destek vektör makinesi, sınıflandırma ve regresyon ağaçları (CART), lojistik regresyon ile yapay sinir ağları yöntemlerini kullanılarak mali tablo dolandırıcılığı tahmin etmiştir. Yücesoy ve ark. [57], çalışmalarında yüksek dereceli tek değişkenli sezgisel bulanık zaman serisi indirgenmiş tahmin modelini tanıtmıştır. Sezgisel zaman serisinin geçmiş bilgisi ile tahmin arasındaki fonksiyonel yapının, yüksek dereceli tek değişkenli sezgisel bulanık zaman serisi indirgenmiş tahmin modeline dayalı karar ağaçlarının torbalanması ile elde edildiği tahmin probleminin çözümü için yeni bir tahmin yöntemi önermişlerdir. Theng ve Boyar [58] son yirmi yılda geliştirilen feature selection algoritmalarını, Büyükkeçeci ve Okur [59] ile Nogueira ve ark. [60] feature selection ve feature selection stabilitesini, Dy ve Brodlay [61] denetimsiz öğrenme için feature selection algoritmalarını detaylı bir şekilde incelemişlerdir. Şenliol ve ark. [62] mRMR algoritmasının kararlılık ölçütlerini ve performansını, Taşçı [63] mRMR algoritmasını kullanarak beyin tümörlerinin sınıflandırılmasını incelemiştir. Robnik-Šikonja ve

Kononenko [64] çalışmalarında ReliefF ve RReliefF algoritmalarının teorik ve ampirik analizini yapmış, Durgabai [65] ise ReliefF algoritmasını kullanarak yapılan öznitelik seçimini araştırmıştır. Şenel ve Alatlı [66] lojistik regresyon analizinin varsayımlarını, sonuçların raporlaştırılmasını ve yorumlanmasını ele almıştır. Zhou ve ark. [67] optimal marjin dağılımına dayalı yeni quadratic yüzey destek vektör makinelerini incelemiştir. Liang ve Liu [68] quadratic kernel ile destek vektör makineleri için verimli özellik ölçeklendirme yöntemlerini ele almıştır. Ghazal ve ark. [69] Fine Gaussian DVM kullanarak Hepatit C evrelemesinin tahmin edilmesi üzerine çalışmışlardır. Uyar ve Uyar [70] çalışmaları ile hipodonti teşhisinde farklı sınıflandırıcılar kullanarak transfer öğrenimini değerlendirmiştir. Shdefat ve ark. [71] insan aktivitesi tanıma sistemlerinde geliştirilmiş DVM ve kNN sınıflandırıcılarının karşılaştırmalı analizini sunmuştur. Maya Gopal ve Bhargavi [72] Boruta algoritmasını kullanarak verim tahmini için öznitelik seçimini ele almıştır. Handhika ve ark. [73] kredi skorlama modelinde Boruta algoritmasını alternatif bir öznitelik seçme yöntemi olarak incelemiştir. Kursa ve ark. [74] Boruta algoritmasının geliştirilmiş bir versiyonunu ve bu algoritmanın performansını araştırmıştır. Demir [75] Parkinson hastalığını konuşma sinyallerinden teşhis etmek için L1-Norm DVM ve ki-kare tabanlı öznitelik seçme algoritmalarını kullanmıştır. Özcan ve Öztürk [76] beyin MR görüntülerinden tümör tespiti için derin öğrenme ve ki-kare test yöntemlerini kullanarak özellik seçimi yapmaktadır. Al ve Özel [77] web sayfası sınıflandırması için filtre tabanlı bir özellik seçme yöntemi olarak ki-kare algoritmasını kullanmıştır. Büyüköztürk [78] kovaryans analizi ve varyans analizini karşılaştırmalı olarak incelemiştir. Demir ve Aslan [79] kNN ve NN algoritmaları ile feature selection tekniklerini kullanarak firewall verilerinin sınıflandırılmasını ele almıştır. Hall [80] rastgele orman ve çok katmanlı algılayıcı tekniklerini içeren korelasyon tabanlı feature selection yöntemini incelemiştir. Sajid ve ark. [81] yüz tanıma için Kruskal-Wallis tabanlı verimli bir öznitelik seçimi yöntemini araştırmıştır. Dass ve ark. [82] gen alt kümelerinin belirlenmesi ve kanser sınıflandırması için Kruskal-Wallis testi ve diğer yöntemleri kullanarak iki aşamalı bir model önermiştir. Yunning Zhong ve ark. [83] EEG patolojisi tespiti için Kruskal-Wallis testi tabanlı bir çerçeve önermiştir. Elssied ve ark. [84] sınıflandırma için F-test kullanarak öznitelik seçimini ele almaktadır. Pintas ve ark. [85] metin sınıflandırması için öznitelik

seçiminden yararlanmıştır. Demir ve Kılıç [86] çalışmalarında denedikleri sınıflandırma tekniklerinden en yüksek performansın rassal orman ve çok katmanlı algılayıcı tekniğini içeren korelasyon tabanlı özellik seçimi yöntemi ve yine aynı oran ile k en yakın komşu tekniğini içeren filtre yöntemi olduğunu açıklamıştır.

# BÖLÜM 3. MATERYAL VE YÖNTEM

Bu bölümde tez kapsamında uygulanan makine öğrenmesi yöntemlerine değinilecektir. Tezde kullanılan verilere ait bağımlı değişkenlerin ikili veri tipine sahip olması ve bu verilerin sınıflandırma yöntemlerine uygun olması nedeniyle bu yöntemler tercih edilmiştir. Veri madenciliğinde sıklıkla kullanılan bu algoritmalardan ve tekniklerden kısaca bahsedilecektir.

## 3.1. Veri Madenciliği

En son bilgi teknolojilerini kullanarak bilimsel araştırma verilerini derlemek ve saklamak giderek kolaylaşmaktadır. Veri öğelerinin sayısındaki artışla gereksiz hale gelecek geleceğe yönelik tahmin yöntemleriyle nasıl başa çıkılacağı sorunu, veri madenciliği tekniklerinin geliştirilmesiyle çözülmüştür. Bu yöntemin amacı, büyük veri setlerinin tahmin edilmesini sağlayacak fonksiyonların belirlenmesidir. Yöntemin hedefi, aynı sorunları ele almak için gelişmiş bilgisayar teknikleri kullanarak klasik istatistiksel tekniklere bir alternatif sağlamaktır. Veri madenciliği, büyük veri havuzlarını faydalı bilgilere dönüştürmekle ilgilenen makine öğreniminin temel alt alanını temsil eder. Veri madenciliği, dönüşüm süreci, veri ön işleme, veri dönüşümü, veri entegrasyonu, veri indirgeme, uygulama ve sunum gibi çeşitli aşamalardan oluşur [87]. Veri madenciliği bağlamında, hedef açıkça tanımlanmışsa, denetimli öğrenme kullanılır; tersine, hedef belirsiz olduğunda denetimsiz öğrenme kullanılır [88].Veri dönüştürme sürecinde yapay zeka, istatistik (veri setlerini sayısal ilişkilerle ilişkilendirme) ve makine öğrenmesinden (tahminler yapmak için veri setlerinden öğrenme) yararlanılmaktadır [89].

## 3.2. Makine Öğrenmesi Yöntemleri

Makine öğrenmesi, tecrübelerden yararlanarak örnek veriyi sınıflandıracak algoritmaları geliştirerek programlayan, istatistiğin gelişmiş teknikleri ile yapay zekanın sezgisel yaklaşımının birleştirilerek geliştirildiği ileri bir halidir. Kısaca makine öğrenmesi istatistik ve yapay zekâ tekniklerinin kesişimi olarak değerlendirilebilir [90, 91].

Makine öğrenmesi modeli kurulurken bağımlı değişken üzerinden öğrenme gerçekleşiyorsa denetimli öğrenme bu model denetimli öğrenme adını almaktadır. Sınıflandırma ve regresyon yöntemleri denetimli öğrenme ile gerçekleşir. En sık kullanılan denetimli öğrenme tekniklerine örnek olarak Destek vektör makineleri (DVM), Yapay sinir ağları (YSA), Karar ağaçları (KA), Rastgele orman (RO), k-en yakın komşu (kNN), Naive Bayes sınıflandırıcısı (NB), Doğrusal regresyon, Lojistik regresyon (LR), Derin öğrenme (DÖ) gibi teknikler verilebilir.

Bağımlı (etiketlenmiş) bir değişken ya da geçmiş deneyimlerle eğitilme imkânı olmayan makine öğrenmesi yöntemlerine denetimsiz öğrenme denir. Benzer veriler yoğunluk tahmini yöntemleriyle değerler alarak oluşturulan modellerle bir araya getirilip gruplandırılır. Denetimsiz öğrenme tekniklerine k-ortalamalar, öz düzenleyici haritalar, hiyerarşik kümeleme, temel bileşenler analizi (TBA) algoritmaları örnek olarak verilebilir.

Yarı denetimli öğrenme ise devam eden örnek olaylarla performansını sürekli geliştiren bir sisteme ulaşmayı hedefler [92].

Makine öğrenmesi modelinin uygulanabilmesi için gerekli bazı aşamaların gerçekleştirilmesi gerekmektedir. Problemin iyi tanımlanması, örnek verinin analizi ve makine öğrenmesi teknikleri için uygun hale getirilmesi (verinin tanınması, gizli bilgilerin bulunması, gürültülü verilerin ayıklanması gibi veri madenciliği tekniklerinin uygulanması), makine öğrenmesinin uygun olduğu algoritmaların seçilip performanslarının değerlendirilmesi adımları gerçekleştirilmelidir.

### 3.2.1. Eğitim ve Test Veri Setlerinin Oluşturulması

Modelin performansı, verinin eğitim seti (train set) ve test seti (test set) olarak ikiye bölünmesi ile ölçülür. Modeli eğitim seti verileri eğitirken modelin performansı test setindeki verilerle ölçülür. Oluşturulan modelin parametreleri içinse seçilmiş olan eğitim setinin içinden doğrulama (validation) seti belirlenir. K eşit parçaya bölünen veri seti için k- katlı çapraz doğrulama tekniği kullanılmıştır. (k-1) sayıda küme eğitim, kalan 1 parça ise test verisi olarak kullanılır. Bu işlem k defa tekrarlanır ve hesaplanan doğruluk değerlerinin ortalaması alınarak, modelin doğruluk değeri hesaplanır. Çalışmalarda genel olarak eğitim ve test veri setleri %20-%80 olarak, çapraz doğrulama katsayısı olan k ise 10 olarak tercih edilir [91].

### 3.2.2. Karışıklık Matrisi (Confusion Matrix)

Makine öğrenmesi ile oluşturulan sınıflandırma modelinin performansı hata matrisi ile değerlendirilir. Karışıklık matrisinde hedef değişkenin tahmini değeri ile gerçek değeri karşılaştırılarak modelin doğruluğu, hassasiyeti, hata oranı, duyarlılığı gibi metrikleri hesaplanır. Bunun için önce Tablo 3.1. ile verilen karışıklık matrisi elde edilir.

Tablo 3.1. Karışıklık Matrisi

| | | Tahmin | |
|---|---|---|---|
| | | Pozitif | Negatif |
| Gerçek | Pozitif | Doğru Pozitif(TP) | Yanlış Negatif (FN) |
| | Negatif | Yanlış Pozitif (FP) | Doğru Negatif (TN) |

TPR (True Positive Rate- Doğru Pozitif Oranı), modelin pozitif sınıfları doğru bir şekilde tahmin etme oranını ifade eder. Duyarlılık (Sensitivity) veya Recall olarak da bilinir.

$$TPR = \frac{TP}{TP+FN} \tag{3.1}$$

FNR (False Negative Rate- Yanlış Negatif Oranı), modelin pozitif sınıfları yanlış bir şekilde negatif olarak tahmin etme oranını ifade eder.

$$FNR = \frac{FN}{TP+FN} \quad (3.2)$$

PPV (Positive Predictive Value- Pozitif Tahmin Değeri), modelin pozitif tahminlerinin doğru olma oranını ifade eder. Precision (Kesinlik) olarak da bilinir.

$$PPV = \frac{TP}{TP+FP} \quad (3.3)$$

FDR (False Discovery Rate-Yanlış Keşif Oranı), modelin pozitif olarak tahmin ettiği örneklerden kaçı yanlış pozitif tahmindir.

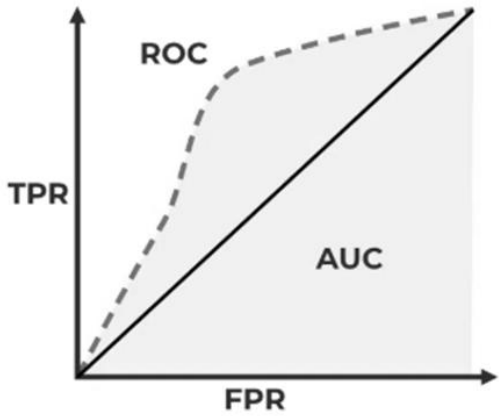
$$FDR = \frac{FP}{TP+FP} \quad (3.4)$$

F1 Ölçüsü, modelin kesinlik ve duyarlılığının harmonik ortalamasıdır.

$$F1 = 2 * \frac{Kesinlik * Duyarlılık}{Kesinlik + Duyarlılık} \quad (3.5)$$

### 3.2.3. ROC Eğrisi ve AUC Değeri

Bir sınıflandırma modelinin performansını değerlendirmek için kullanılan bir grafiktir. Bu grafik, modelin farklı eşik değerlerinde ürettiği Doğru Pozitif Oranı (TPR) ve Yanlış Pozitif Oranı (FPR) arasındaki ilişkiyi gösterir. ROC (Receiver Operating Characteristic Curve) eğrisi, bir modelin doğruluk ve hata arasındaki dengesini gösterir (Şekil 3.1). Eğrinin ideal şekli, sol üst köşeye yakın bir çizgi olmalıdır; bu, modelin TPR değerini artırırken aynı zamanda FPR değerini ise mümkün olduğunca düşük tuttuğunu ifade eder.




Şekil 3.1. Roc Eğrisi ve AUC Değeri

AUC (Area Under the Curve) değeri ROC eğrisinin altındaki alanı ifade eder ve modelin genel performansını özetler. AUC değeri 0.5 ile 1 arasında değişir:

- 1.0: Mükemmel bir model (hiçbir hata yapmaz)
- 0.5: Rastgele tahmin yapan bir model (başarısız model)

AUC, bir modelin pozitif örnekleri negatiflerden ayırma yeteneğini ölçer. Daha yüksek AUC değeri, daha iyi ayrım gücü anlamına gelir.

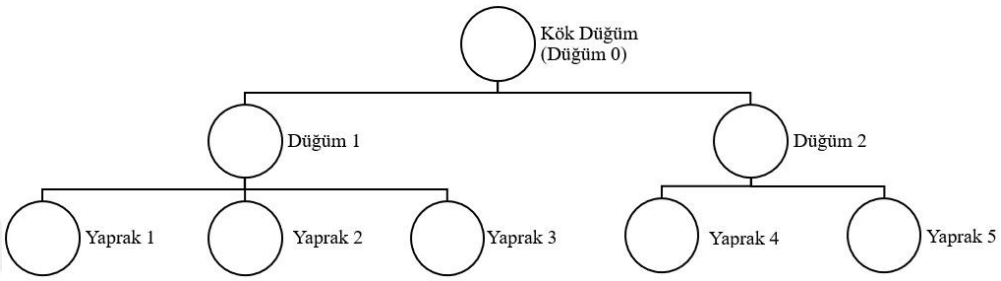
### 3.3. Ağaç Tabanlı Yöntemler

Ağaç tabanlı yaklaşımlar, bir veya daha fazla karar ağacından tahmin yapmak için bir dizi koşullu ifadeden yararlanır. Tüm ağaç tabanlı modeller, sayısal değerleri öngörmek için regresyon veya kategorik değerleri belirlemek için sınıflandırma amacıyla kullanılabilir. Tez çalışması kapsamında yer aldıkları için burada karar ağaçları ve rastgele orman algoritmalarından bahsedilecektir.

#### 3.3.1. Karar Ağaçları

Karar ağacı yapısı, yapım kolaylığı ve anlaşılabilirliği nedeniyle sınıflandırma ve regresyon modelleri için sıklıkla tercih edilen bir yöntemdir [57]. Normal dağılım ve homojen varyans varsayımı yapan parametrik olmayan istatistiksel veriler için karar

ağaçlarının yapısı "if-then" koşullu ilişkilerine dayanmaktadır. Bağımlı değişken ile bağımsız değişkenler arasındaki ilişki, bağımsız değişkenlerin ilişki seviyelerine göre sınıflandırılır. Sınıflandırma, her bir değişkenin diyagramdaki bir düğüm üzerinde gösterilmesiyle oluşturulur (Şekil 3.2). Ağaç, tüm örnekleri temsil eden bir kök düğümle başlar ve daha sonra ağacı dallara ayırarak verileri alt gruplara bölmeye devam eder [93].




Şekil 3.2. Karar Ağacı Yapısı

Karar ağaçlarında en sık kullanılan tekniklerden biri CHAID algoritmasıdır. CHAID tekniği, özellikle kategorik bağımlı değişkenler için tasarlanmış olan AID tekniğinin geliştirilmiş bir versiyonudur [15, 34]. CHAID algoritmasında bağımlı değişken nicel veya kategorik, bağımsız değişken ise kategorik olmalıdır. CHAID algoritması bir düğümün iki veya daha fazla dala bölünmesine izin verir. Bağımlı değişken bir ağaç yapısı kullanılarak regresyona tabi tutulabilir ya da sınıflandırılabilir [94]. Bölme kriteri olarak kullanılan F testi veya ki-kare test istatistiği, mevcut veri kümesini iteratif bir şekilde alt gruplara ayırır ve bağımlı değişkeni etkileyen bağımsız değişkenleri etki büyüklüklerine göre ağaç benzeri bir yapıda sıralar [93]. Ki-kare analizi, iki kategorik değişken arasındaki ilişkinin anlamlılığını hesaplamak için kullanılan parametrik olmayan bir yöntemdir. İstatistiksel tekniklerin uygulanması, özellikle de iki kategori arasındaki ilişkinin incelenmesine olanak tanıyan bir olasılık tablosunun oluşturulması yoluyla elde edilir. $r \times c$ tipinde bir çapraz tablo için, $i = 1,2,\ldots,r$ satır indisleri ve $j = 1,2,\ldots,c$ sütun indisleri olsun. $i$-inci sıra ve $j$-inci sütuna ait gözlem değeri $G_{ij}$, beklenen değer ise $B_{ij}$ ile gösterilmek üzere $\chi^2$ test istatistiği (3.6) formülü kullanılarak elde edilir:

$$\chi^2{}_{test} = \frac{\sum_{j=1}^{c}\sum_{i=1}^{r}(G_{ij}-B_{ij})^2}{B_{ij}} \quad (3.6)$$

Eğer $\chi^2{}_{test} > \chi^2{}_{(r-1)(c-1)}$ oluyorsa beklenen ve gözlenen değerler arasındaki farkın anlamlı olmayıp, iki değişken arasında istatistiksel olarak anlamlı bir ilişki olduğu sonucuna varılır. Prosedürler, her bir bağımsız değişken ile bağımlı değişken arasındaki ilişkinin belirlendiği bir şekilde yürütülür ve süreç tüm gözlemler homojen hale gelene kadar ağaç dallanma kriterlerine göre devam eder [95].

*l* adet bağımsız değişken ($k = 1, 2, \dots, l$) ve bir bağımlı değişken ($Y$) tanımlansın. *k.* bağımsız değişkenin $r$ ($i = 1,2, \dots, r$), bağımlı değişkenin ise $c$ ($j = 1,2, \dots, c$) alt kategorisi olsun.

Adım 1. Eğer $X_k$ bağımsız değişkeni 2 alt kategoriye sahipse (*2xc* boyutlu ise) prosedür Adım 5'e ilerler. Aksi takdirde Adım 2'ye geçilir.

Adım 2. Amaç, tüm olası kombinasyonları içeren iki boyutlu çapraz tabloları oluşturarak $\chi^2{}_{test}$ değerini hesaplamaktır. $r > 2$ olan $X_k$ 'nın alt kategorilerinin ikili kombinasyonları düşünülerek $C(r, 2) = \frac{r!}{2!(r-2)!}$ tane kontenjans tablosu elde edilir.

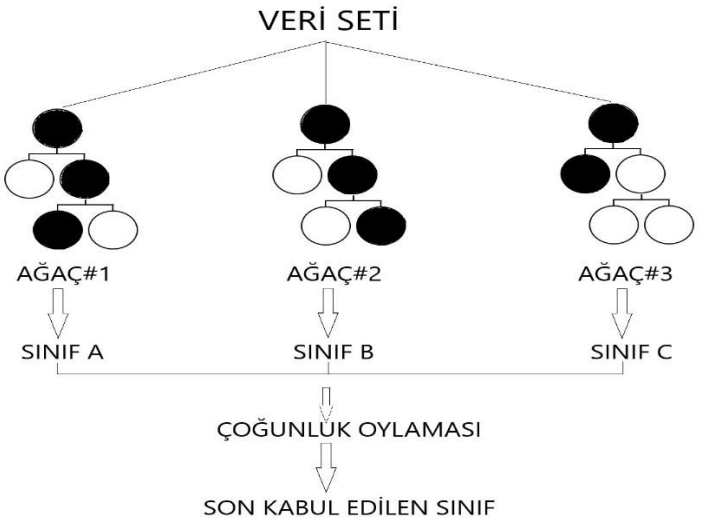
Adım 3. $X_k$ ve $Y$ ile elde edilen $\chi^2{}_{test} < \chi^2{}_{(1)(c-1)}$ şartını sağlayan en düşük $\chi^2{}_{test}$ istatistiği seçilir. $X_k$ değişkeninin iki alt kategorisi birleşik bir kategoride birleştirilir. r değeri r-1 olarak güncellenir ve süreç Adım 1'e geri döner. Bu adım, oluşturulan tüm ikili çapraz tablolar için $\chi^2{}_{test} > \chi^2{}_{(1)(c-1)}$ şartı gerçeklenene kadar yinelenir. Bu koşul sağlandığında, süreç Adım 4'e ilerler. Önceki 1-4 adımları tüm bağımsız değişkenlere $(X_1, X_2, \dots, X_l)$ uygulanır.

Adım 4. $(X_1, Y), \dots, (X_k, Y), \dots, (X_l, Y)$ değişken çiftleri için $2xc$ boyutlu çapraz tablolarının $\chi^2{}_{test}$değerleri hesaplanır. En büyük $\chi^2{}_{test}$ değerine sahip olan $X_k$ değişkeninin alt kategorileri aracılığı ile dallanma ve ardından alt düğümler elde edilir. Algoritmanın adımları her bir alt düğüm için tekrarlanır. Bir düğümdeki tüm gözlemler

bağımlı değişkenin aynı kategorisinde bulunuyorsa veya $(X_k, Y)$ değişken çiftine ait oluşturulan tüm mümkün $2xc$ çapraz tablosunda $\chi^2{}_{test} < \chi^2{}_{(1)(c-1)}$ koşulu sağlanıyorsa bu düğüme terminal düğümü denir. Terminal düğümde yapılan gözlemler homojendir ve farklı kategorilerde sınıflandırılamaz. İşlem ağaçtaki tüm düğümler terminal olana kadar devam eder.

### 3.3.2. Rastgele Orman (Random Forest)

Random Forest (Rastgele Orman Algoritması), birden çok karar ağacı üzerinden her bir karar ağacını farklı bir öğrenme örneği seti üzerinde eğiterek çeşitli modeller üretip, sınıflandırma oluşturmayı sağlayan, çeşitli veri kümeleri üzerinde özel işlemler ile farklı çözümlemeler için yapılandırılmış denetimli/gözetimli öğrenme algoritmalarından regresyon ve diğer analizler için eğitim aşamasında çok sayıda karar ağacı oluşturarak problemin tipine göre sınıf/sonuç tahmini gerçekleştiren toplu öğrenme yöntemidir. Rastgele orman algoritması örnek şeması Şekil 3.3. ile verilmiştir:




Şekil 3.3. Rastgele Orman Örnek Şeması

Rastgele orman algoritmasında

- Analiz edilecek veri seti hazırlanır,

- Algoritma her bir bölünmede seçilen bağımsız değişken sayısına göre oluşturabilecek bütün karar ağaçlarından her biri için tahmin değerleri oluşturulur,
- Tahmin sonucu oluşan her değer için bağımsız değişkenin önem derecesine göre sonuçlar ağırlıklandırılarak oylama gerçekleştirilir,
- Son olarak algoritmada en çok ağırlıklı bulunan değişken(ler) sonucu oluşturur.

Algoritma kurulurken öncelikle eğitilmiş olan k tane karar ağacı (3.7) denklemindeki kriterlerle oluşturularak biraraya getirilir:

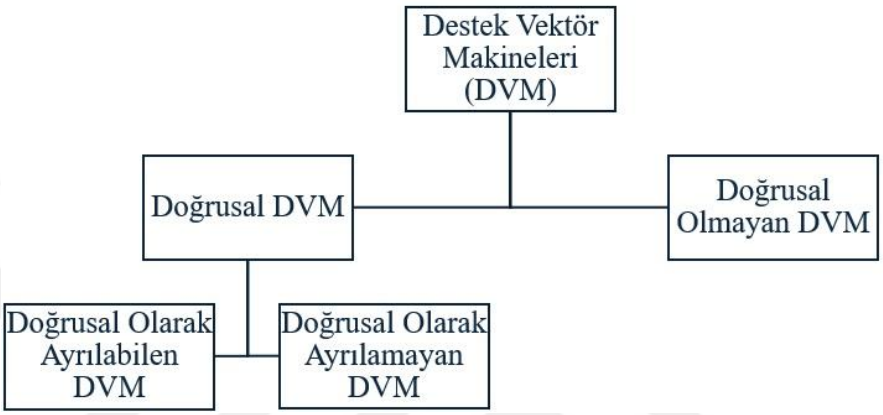
$$H(X, \theta_j) = \sum_{i=0}^{k} h_i(x, \theta_j)\ , (j = 1,2, \dots, m) \qquad (3.7)$$

Denklem (3.7) ile tanımlanan $H(X,\ \theta_j)$ bir meta karar ağacı sınıflandırıcısıdır . $x$, eğitim veri kümesinin girdi özelliği vektörünü, $\theta_j$ ağacın büyüme sürecini belirleyen bağımsız ve aynı şekilde dağıtılmış rastgele bir vektörü temsil etmektedir [102]. Bu kriterle oluşturulan k karar ağacı, test veri kümesinin tüm örneklerini tahmin etmektedir. Elde edilen bu tahmin sonuçlarına göre sınıflandırma gerçekleştirilmektedir [103]. Klasik karar ağaçları tekniğinde aşırı öğrenme (overfitting) bir dezavantaj oluştururken RF algoritmasında veri setini ve öznitelikleri parçalara bölüp çok sayıda ağaç üzerinde işlem gerçekleştiği için bu problemi ortadan kaldırmaktadır. için hem veri setini hem de öznitelikleri çok sayıda parçaya bölerek birden fazla ağaçta işlem gerçekleştirmektedir [104]. Bu çalışma prensibi ile RF, özellikleri büyük veri tabanlarında bile eksiksiz işlediği ve en çok etkiye sahip özellikleri ön plana çıkarmada son derece yüksek bir performansa sahip olduğu için sınıflandırma yaparken düşük hatalarla yüksek doğruluk oranını yakalayabilmektedir [105].

### 3.4. Destek Vektör Makineleri

Destek Vektör Makinesi (DVM), denetimli bir öğrenme modelini temsil eder. Giriş verileri belirli bir sınıfa göre etiketlenir. DVM, girdi olarak sağlanan iki veri sınıfı

arasında ayrım yapmak için n boyutlu bir hiper düzlem kullanır. Hiper düzlemler aynı zamanda karar sınırları olarak da adlandırılır. Karar sınırı, her bir sınıfın en yakın veri noktalarından mümkün olduğunca uzak olacak şekilde oluşturulur. Hiper düzlemi tanımlayan veri noktaları "destek vektörleri" olarak adlandırılır. Şekil 3.4 ile belirtildiği üzere destek vektör makineleri genel olarak doğrusal veya doğrusal olmayan DVM şeklinde ikiye ayrılmaktadır.




Şekil 3.4 Doğrusallık Yapısına göre DVM

Doğrusal olmayan modellerde, iki sınıf arasındaki mesafe, bu özel amaca sahip matematiksel bir yapı olan bir kernel (çekirdek fonksiyonu) kullanılarak hesaplanır. Kernel, yüksek boyutlu ve doğrusal olmayan modellerin oluşturulmasını sağlayan bir fonksiyondur. Doğrusal olmayan problemler bağlamında, çekirdek fonksiyonu işlenmemiş veriyi ek boyutlarla güçlendirmek için kullanılabilir ve böylece işlenmemiş veriyi daha yüksek boyutlu bir uzayda doğrusal bir probleme dönüştürür. Kernel fonksiyonu, belirli hesaplamaların hızlı bir şekilde tamamlanmasını kolaylaştırmak için tasarlanmıştır.

Doğrusal DVM' nin karar düzlemi, sınıflar arasında ayrım yapmada oldukça etkilidir. Başka bir deyişle, iki sınıflı etiketli veri setleri için iki sınıf arasında belirgin bir sınır tanımlanabilir. Denklem (3.8)'de belirtildiği gibi, etiketleri bilinen (denetimli öğrenme) bir veri seti için, $x_i$ girdi vektörlerini, $y_i$ ise +1 veya -1 değerlerini alabilen sınıf etiketlerini temsil eder. n değeri, veri setinde bulunan toplam özellik sayısını gösterir.

$$X = \{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n), x_i \in \mathbb{R}^d, y_i \in \{-1, +1\}\} \quad (3.8)$$

*f(x)*, hiperdüzlemle bağıntılı bir diskriminant fonksiyonunu temsil edip, karar fonksiyonu *sgn(f(x))* ile tanımlansın. (3.9) denklemi ile formüle edilen fonksiyonda *w* optimum hiperdüzlemi tanımlayan vektörü, *b* ise skaleri temsil etmektedir.

$$f(x) = w^T x_i + b,\ w \in \mathbb{R}^d \text{ ve } b \in \mathbb{R} \quad (3.9)$$

Optimum hiperdüzlemin (3.10) ile verilen şartları sağlaması beklenir:

$$\begin{cases} w^T x_i + b \leq -1 \ eğer \ y_i = -1 \\ w^T x_i + b \geq -1 \ eğer \ y_i = +1 \end{cases} \quad (3.10)$$

$H_1$: $wx_1^T + b = +1$ ve $H_2$: $wx_2^T + b = -1$ ile formülize edilen $H_1$ ve $H_2$ hiperdüzlemleri arasındaki mesafeyi temsil eden bölgeye sınır bandı denir. (3.10) ile verilen denklemler üzerinden bu mesafe $\frac{2}{\|w\|}$ olarak hesaplanır [96]. $H_1$ ve $H_2$ düzlemlerine eşit uzaklıklı bir $H_0$ medyanı olduğunu kabul edersek (3.11) denklemi ile verilen $H_0$ düzlemi:

$$H_0 : wx_0^T + b = 0 \quad (3.11)$$

$H_0$ ve $H_1$ düzlemleri arasındaki uzaklık $d^{\text{positive}}$, benzer şekilde $H_0$ ve $H_2$ düzlemleri arasındaki uzaklık $d^{\text{negative}}$ ile temsil edilir ve bu mesafe değerleri $\frac{1}{\|w\|}$ olarak hesaplanır.

Bir DVM modelini eğitmenin temel amacı, w ve b değerlerini hesaplamaktır. Bu, yüksek boyutlu verilerin en uygun şekilde bölümlenmesini ve böylece marjın en üst düzeye çıkarılmasını sağlar [89].

İki sınıfın doğrusal olarak ayrılabilmesi durumunda, daha önce açıklandığı gibi, hiper düzlem üzerindeki en yakın destek noktaları, destek vektörüne dayalı mesafe ölçümleri

kullanılarak belirlenir. Hiper düzlemi belirlemek için maksimum mesafe kullanılır ve bu mesafe daha sonra iki sınıfı ayırmak için kullanılır. İki sınıfın doğrusal olarak ayrılamaması durumunda, destek vektörleri potansiyel sınıflandırma hatasını dikkate alarak maksimum hiper düzlemi belirlemeye çalışır. Verilerin yukarıda bahsedilen doğrusal olmama durumu, hata terimlerini temsil eden değişkenlerin dahil edilmesiyle çözülebilir. Doğrusal olmayan destek vektör makinelerinde problem, doğrusal olmayan çekirdek fonksiyonlar kullanılarak verilerin doğrusal bir formata dönüştürülmesiyle çözülebilir. Literatürde karşılaşılan en iyi bilinen çekirdek fonksiyonlar doğrusal, radyal, polinom ve sigmoid çekirdek fonksiyonlarıdır [97].

- Lineer çekirdek fonksiyonunun kullanım alanları yüksek boyutlu ve doğrusal olarak ayrılabilir veri kümeleridir. $K(x_i, x_j) = x_i . x_j$ dönüşümü ile tanımlı doğrusal çekirdek fonksiyonunun parametresi yanlızca $C$ (düzenleme parametresi) dir.
- Radyal tabanlı çekirdek fonksiyonu karmaşık ayrımları yakalayabildiği için genellikle en iyi performansı verir. Kullanım alanları karmaşık sınıflandırma problemleri, görüntü işleme ve biyoinformatik olan radyal çekirdeği $K(x_i, x_j) = e^{(-\gamma \|x_i - x_j\|^2)}$ dönüşümü ile tanımlıdır. Burada $\gamma$ parametresi etki alanını belirler. Küçük γ, daha geniş bir etki alanı oluşturur. Lineer olarak ayrılmayan verilerde daha iyi performans gösterir. Hesaplama maliyeti lineere göre daha yüksektir.
- Polinomial çekirdeği, orta düzeyde karmaşık olan veri kümelerinde, görüntü ve el yazısı uygulamalarında daha iyi performans gösterir. $K(x_i, x_j) = (x_i . x_j + c)^d$ ile tanımlı polinomial çekirdek fonksiyonunda yer alan $d$ polinomun derecesini belirler. Daha büyük $d$ değeri, daha karmaşık karar sınırları oluşturur. $c$ (sabit terim) genellikle 1 olarak alınır. Düşük dereceli polinomlar genel olarak daha iyi geneller.
- Sigmoid çekirdek fonksiyonu genellikle sinir ağları ile benzer yapıda çalışan problemler, biyolojik sinyaller ve nöron aktivitesi modelleme konularında iyi performans göstermektedir. $K(x_i, x_j) = tanh(\alpha x_i . x_j + c)$ ile tanımlı sigmoid

çekirdeğinde yer alan $\alpha$ (ölçekleme) ve $c$ (kaydırma) parametreleri, ağın esnekliğini belirler.

Tez çalışması kapsamında yer aldıkları için burada kısaca Kuadratik (Karesel) DVM ve Fine Gaussian DVM algoritmalarından bahsedilecektir.

### 3.4.1. Kuadratik (Karesel) DVM Algoritması

Kuadratik DVM'ler, veriyi farklı sınıflara ayıran bir hiper düzlem bulmayı amaçlayan güçlü bir sınıflandırma yöntemidir. Doğrusal olmayan ayrım gerektiren durumlarda kuadratik (ikinci dereceden) bir karar sınırı kullanır. Bu, genellikle veri setinde sınıflar arasında doğrusal bir ayrım yapılamadığında tercih edilir. Kuadratik DVM, veriyi sınıflara ayırmak için doğrusal olmayan (parabolik veya kuadratik) bir hiper düzlem oluşturur. Bu karar sınırı, doğrusal DVM' lerin yetersiz kaldığı durumlarda daha etkili ayrım yapabilir. Kuadratik DVM, genellikle doğrusal olmayan bir kernel fonksiyonu (genellikle Polynomial Kernel) kullanarak gerçekleştirilir.

### 3.4.2. Fine Gaussian DVM Algoritması

Fine Gaussian terimi, Gaussian kernel (RBF- Radial Basis Function) kullanılarak gerçekleştirilen bir DVM modelini ve bu modelin, küçük bir sigma (bant genişliği) değerine sahip olduğunu ifade eder. Küçük sigma değeri, daha karmaşık ve detaylı karar sınırları öğrenmek için modelin daha hassas bir şekilde ayarlandığı anlamına gelir. Gaussian kernel, her veri noktası için bir "etki alanı" tanımlar ve doğrusal olmayan ayrımları mümkün kılar.

Veriler, Gaussian kernel fonksiyonu kullanılarak yüksek boyutlu bir uzaya dönüştürülür. Verileri sınıflar arasında en iyi ayıran karar sınırı öğrenilir. Bu sınır, doğrusal olmayan ve genellikle eğrisel bir yapıdadır. Hiperparametreler (sigma ve düzenleme parametresi C) optimize edilerek sınıflandırma doğruluğu artırılır.

## 3.5. Lojistik Regresyon

Lojistik regresyon, iki veya daha fazla kategoriden oluşan bir bağımlı değişkene, her bir bağımsız değişkenin etkisini araştırmak için kullanılan bir sınıflandırma yöntemidir. Kategorik değişkenlerin bir ölçeklendirme sistemine göre sınıflandırıldığı durumlarda, lojistik regresyon analizi uygun bir analitik tekniği temsil etmektedir [98]. Bu, yanıt değişkeninin beklenen değerlerinin bağımsız değişkenlere göre olasılıklı bir şekilde elde edildiği bir regresyon tekniğidir. Normal dağılım varsayımı, kalıcılık ön varsayımını gerektirmez. Amaç, bağımsız değişkenlerin bağımlı değişken üzerindeki etkilerini belirlemektir [99]. Bu modelde bağımlı değişkenler kategorik olup kesikli, bağımsız değişkenler ise sürekli değerlerle açıklandığı için modelin $0 \leq E(Y_i/X_i) \leq 1$ şartını sağlaması gerekir. Birikimli lojistik dağılım fonksiyonu, $Z_i = b_1 + b_2X_i$ olmak üzere (3.12) denklemindeki gibi tanımlanır:

$$P_i = E(Y = 1/X) = \frac{1}{1+e^{-Z_i}} = \frac{1}{1+e^{-(b_1+b_2X_i)}} \tag{3.12}$$

Denklemde yer alan $P_i$, X bağımsız değişkeni ile ilgili bilgi veri olduğunda i' inci değişkenin belirli bir seçimi yapma olasılığını göstermektedir ($0 \leq P_i \leq 1, -\infty < Z_i < \infty$). $P_i$ ve $Z_i$ arasında doğrusal bir ilişki olamayacağı aşikâr olduğu için yöntemin parametre tahminlerinde En Küçük Kareler Yöntemi (EKK) kullanılamamaktadır. Lojistik regresyon analizinde bir olayın gerçekleşme olasılığının gerçekleşmeme olasılığına oranına bahis oranı (odds ratio) denir ve (3.13) denklemi ile hesaplanır.

$$Odds\ Ratio = \frac{P_i}{1-P_i} = \frac{1}{1+e^{-Z_i}} . \frac{1+e^{-Z_i}}{e^{-Z_i}} = e^{Z_i} \tag{3.13}$$

Bahis oranı sayesinde lojistik regresyon fonksiyonu, doğrusal regresyon için uygun hale getirilebilir. (3.13) denkleminin her iki tarafının doğal logaritması alınırsa doğrusal yapılı (3.14) denklemi elde edilir.

$$g(x) = \ln\left(\frac{P_i}{1-P_i}\right) = lne^{Z_i} = Z_i = b_1 + b_2X_i \tag{3.14}$$

Değişkenler vektörü $x' = [x_1, x_2, \dots, x_p]$ olan lojistik regresyon denklemi $g(x) = b_0 + b_1x_1 + b_2x_2 + \cdots + b_px_p$ olarak tanımlanmışsa $g(x)$, çok değişkenli bir LR fonksiyonudur.

Lojistik regresyon üç ana kategoride sınıflandırılabilir.

### 3.5.1. İkili Lojistik Regresyon

İki kategoriye ayrılan bağımlı değişkenleri ve sürekli ya da kategorik bir özelliğe sahip bağımsız değişkenleri analiz etmek için kullanılan istatistiksel bir yöntemdir. Bağımlı değişken, incelenen sonucun bir ölçüsüdür ve belirli bir bağlamda 1, diğer senaryolarda ise 0 değerini alır. Bağımsız değişkenler, bu tür bir modelleme yoluyla belirlenen gruplara dahil olan birimlerle ilgilidir. Bir birimin ilk kategoriye ait olma olasılığı (3.15) ile verilen denklem kullanılarak hesaplanır:

$$\mathrm{P}(Y_i = 1) = \frac{e^{(b_0 + b_1x_{i1} + \dots + b_kx_{ik})}}{1 + e^{(b_0 + b_1x_{i1} + \dots + b_kx_{ik})}} \qquad (3.15)$$

### 3.5.2. Sıralı Lojistik Regresyon

Bağımlı değişken ikiden fazla kategori içerdiğinde ve bu kategoriler en küçükten en büyüğe doğru sıralandığında kullanılır. İçinde $Y_i$ kategorisi, olasılık değeri *r* ile, ilgili kategori için kenar değeri *i* ile, regresyon katsayıları *b* ile ve yukarıda belirtilen kategoriler için ölçeklendirme parametreleri hakkında fikir veren ölçeklendirme parametreleri vektörü *Z* ile gösterilir. Bu olasılık değeri (3.16) ile verilen denklem kullanılarak hesaplanır:

$$Link(Y_i) = \frac{r_i[b_0 + b_1X_1 + \dots + b_kX_k]}{e^{(\phi_0 + \phi_1Z_1 + \dots + \phi_tZ_t)}} \qquad (3.16)$$

### 3.5.3. Multinomial lojistik regresyon

Bağımlı değişkenin iki veya daha fazla kategorik sınıflandırma ölçeğine sahip olduğu durumlarda kullanılmalıdır. Örneğin, üç farklı akademik programa kayıtlı öğrencilerden oluşan bir bağımlı değişken varsa, multinomial lojistik regresyon analizi kullanılarak analiz edilebilir [100]. Bu olasılık değeri (3.17) ile verilen denklem kullanılarak hesaplanır:

$$\mathrm{P}(Y_i = n) = \frac{e^{(b_{n0}+bn_1X_{i1}+\ldots+b_{nk}X_{ik})}}{1+\sum_{n=1}^{M-1} e^{(b_0+b_1X_{i1}+\ldots+b_kX_{ik})}} \tag{3.17}$$

## 3.6. Öznitelik Seçimi (Feature Selection)

Makine öğrenmesinde özellik seçimi, bir modeli kurarken ilgili özellikler arasından (değişkenler, öngörücüler) bir alt küme seçme sürecidir. Amaç, makine öğrenmesi modelini oluştururken tahmin değişkenine en çok katkıda bulunan bağımsız değişkenleri seçerek modelin tahmin gücünü artırmaktır. Öznitelik seçimi için manuel veya otomatik pek çok yöntem bulunmaktadır. Tez çalışması kapsamında yer aldıkları için burada kısaca mRMR, Ki-Kare, ReliefF, ANOVA ve Kruskal-Wallis algoritmalarından bahsedilecektir.

### 3.6.1. mRMR Tabanlı Öznitelik Seçimi

Minimum Redundancy Maximum Relevance (mRMR) algoritması, sınıflandırma ve regresyon görevlerinde kullanılan popüler bir özellik seçme yöntemidir. Bu algoritma, hedef değişkenle olan ilişkiyi (relevance) en üst düzeye çıkarırken, özellikler arasındaki fazlalığı (redundancy) en aza indirmeyi amaçlayarak özellikleri sıralar. $mRMR = Relevance - Redundancy$ skorunu maksimize eden özellik seçilir. Özellikler tek tek eklenir ve belirli bir sayıya ulaşıldığında veya tüm özellikler sıralandığında işlem sonlandırılır.

Algoritma, her bir özniteliği ve sınıf etiketleri vektörünü ayrık bir rastgele değişken olarak ele alır ve iki öznitelik arasındaki ya da bir öznitelikle sınıf etiketleri vektörü arasındaki benzerlik düzeyini değerlendirmek için ortak bilgiden (mutual information), $I(X, Y)$, yararlanır.

$$I(X,Y) = \sum_{y\epsilon Y} \sum_{x\epsilon X} p(x,y)\log\left(\frac{p(x,y)}{p_1(x)p_2(y)}\right) \quad (3.18)$$

(3.18) denkleminde $p(x,y)$ $X$ ve $Y$ rastgele değişkenlerinin bileşik olasılık dağılım fonksiyonunu, $p_1(x)$ ve $p_2(y)$ ise $X$ ve $Y$ rastgele değişkenlerinin marjinal olasılık dağılım fonksiyonlarını göstermektedir. $I(X,Y)$ ortak bilgisi, $X$ ve $Y$ rastgele değişkenlerinin tam bağımsız olması durumunda 0 değerini alır ($I(X,Y) \geq 0, I(X,Y) = I(Y,X)$).

Öznitelikler, $n$ tane gözlemin oluşturduğu $f_i = [f_i{}^1, f_i{}^2, f_i{}^3, \dots, f_i{}^n]$ vektörü ile tanımlansın. $f_i$ bir öznitelik vektörünün reel değerini barındıran bir vektör, $F_i$ ayrık rassal değişkeninin bir örneği ise, $i$ ve $j$ öznitelikleri arasındaki ortak bilgi $I(F_i, F_j)$, bu ayrık rassal değişkenler üzerinden hesaplanacaktır. $i, j = 1,2, \dots, d$ olmak üzere $d$, veri kümesinin öz nitelik sayısını göstermektedir. Ortak bilgi kavramı aynı zamanda herhangi bir öznitelik ile sınıf etiketleri vektörü, $h(h = [h^1, h^2, h^3, \dots, h^n])$ ve ona tekabül eden ayrık rassal değişkeni, $H$, arasındaki benzerliğin, $I(H, F_i)$, ölçülmesinde de kullanılır.

$S$ seçilmek istenen öznitelik kümesi, $|S|$ise bu kümenin eleman sayısını belirtsin. Burada $S$'nin seçilebilecek en iyi öznitelik kümesi olduğunu göstermek için maksimum ilişki ve minimum artıklık koşullarının sağlanması beklenir. Maksimum ilişki (maximum relevance) (3.19) denklemi ile hesaplanır:

$$\max W, W = \frac{1}{|S|}\sum_{F_i\in S} I(H, F_i) \quad (3.19)$$

Minimum artıklık (minimum redundancy) (3.20) denklemi ile hesaplanır:

$$\min V, V = \frac{1}{|S|^2}\sum_{F_i,F_j\in S} I(F_i, F_j) \quad (3.20)$$

mRMR algoritması yukarda bahsedilen iki koşulun $\max(V - W)$, $\max(V/W)$ kombinasyonu altında birleştirilmesi ile ortaya çıkmıştır. mRMR algoritmasında önce

ilk öznitelik (3.19) eşitliği ile seçilir. Bundan sonraki her adımda (3.21) ve (3.22) koşullarını gerçekleyen $F_i$ özniteliği seçilir ve seçilen öznitelikler *S* kümesinde tutulur. *S* tüm öznitelik kümesini, $\Omega_S$ ise seçilmiş öznitelikler dışında kalan tüm öznitelikleri ifade etmektedir *($\Omega_S = \Omega - S$).*

$$\max_{F_i \epsilon \Omega_S} [I(H, F_i) - \frac{1}{|S|} \sum_{F_j \in S} I(F_i, F_j)] \quad (3.21)$$

$$\max_{F_i \epsilon \Omega_S} \{I(H, F_i) / [\frac{1}{|S|} \sum_{F_j \in S} I(F_i, F_j)]\} \quad (3.22)$$

İlk öznitelik seçildikten sonra (3.21) ya da (3.22)' yi sağlayan öznitelik *S* kümesine dahil edilir ve istenilen öznitelik sayısına ulaşıldığında algoritma sonlandırılır [106].

### 3.6.3. Ki-Kare İstatistiği ile Öznitelik Seçimi

Öznitelik seçiminde sıklıkla başvurulan tekniklerden biri de Ki-Kare istatistiğidir. Bu yöntemle değişkenlerin veri setini tanımlamak için uygun olup olmadığı belirlenir. İlk aşamada gözlenen değerlerin gerçek sınıflara göre Ki-Kare ($\chi^2$) test istatistiği hesaplanır. Sıfır ile pozitif sonsuzluk arasında değer alan $\chi^2$ değeri, gözlenen ve beklenen değer arasındaki uyum arttıkça sıfır değerine yaklaşır. $\chi^2$ büyüdükçe gözlenen ve beklenen değerler arasındaki uyumsuzluğa işaret edilmektedir. Bunun için ikinci aşamada Ki-Kare dağılımında önemlilik seviyesi ve serbestlik derecesi ile belirlenen eşik değeri, $\chi^2$ değeri karşılaştırılır. Önemlilik seviyesi, araştırmacı tarafından belirlenen yüzde değeri, serbestlik derecesi ise öznitelik sayısının bir eksiğidir. $\chi^2$ (3.23) denklemi ile hesaplanır:

$$\chi^2 = \sum_{i=1}^{n} \frac{(o_i - e_i)^2}{e_i} \quad (3.23)$$

Burada *n* öznitelik sayısını, $o_i$, i'inci öznitelik için gözlenen frekans değerini, $e_i$ ise i'inci öznitelik için beklenen frekans değerini temsil etmektedir [107].

### 3.6.4. ReliefF Algoritması ile Öznitelik Seçimi

ReliefF algoritması çoklu sınıflara sahip veri setleri için, öznitelikleri aralarındaki ilişkiye göre ağırlıklandıran bir öznitelik seçim yöntemidir. Algoritmanın başlangıcında tüm özniteliklerin ağırlıkları 0 olarak belirlenir. Daha sonra her bir aşamada rastgele olarak seçtiği veri ile aynı sınıfa üye en yakın k (k değeri sınıf sayısının bir eksiğidir) tane veri bularak her bir farklı sınıfa ait en yakın veriler bulunur. Sonrasında her bir özelliğe ait ağırlıklar bu veriler kullanılarak güncellenir. Son aşamada belirlenen koşula uymayan özellikler veri setinden ayrılarak yeni veri seti elde edilir. ReliefF algoritması (3.24), ikili değerler için uzaklık hesaplanması (3.25), devamlı değerler için uzaklık hesaplanması ise (3.26) ile formülize edilmiştir.

$$\mathrm{W}(x^a) = \mathrm{W}(x^a) - \frac{\sum_{j=1}^{k} uzaklık(A,R_i,H_j)}{m \times k} + \frac{\sum_{C \neq sınıf(R_i)} \left[ \frac{P(C)}{1-P(sınıf(R_i))} \times \sum_{j=1}^{k} uzaklık(A,R_i,M_j) \right]}{m \times k} \tag{3.24}$$

$$uzaklık(A, I_1, I_2) = \begin{cases} 0, I_1 = I_2 \\ 1, I_1 \neq I_2 \end{cases} \tag{3.25}$$

$$uzaklık(A, I_1, I_2) = |I_1 - I_2| \times \frac{1}{\max(A) - \min(A)} \tag{3.26}$$

Yukarıda verilen denklemlerde $\mathrm{W}(x^a)$, a' ıncı özniteliğin ağırlığını, k sınıf sayısının 1 eksiğini, m döngü sayısını, $R_i$ i'inci döngüde seçilen veriyi, $H_J$ seçilen veri ile aynı sınıfa ait j' inci yakın veriyi, $M_j$ j' inci sınıfa ait seçilen veriye en yakın veriyi temsil etmektedir [107].

### 3.6.5. ANOVA Tabanlı Öznitelik Seçimi

ANOVA, farklı grupların ortalamalarını istatistiksel olarak değerlendirmek için F-testini kullanarak her özelliğin etiketine göre değerini belirleyebilir. Daha sonra her özellik puanlanır ve etiketle en alakalı puanı hangisinin aldığını görmek için sıralanır. ANOVA algoritması ile özellik sayısı ve bir "*F* oranı" elde edilir. *F* oranı ne kadar yüksekse, sınıf o kadar ayrıdır. *F* oranı, sınıflar arası varyansın sınıf içi varyansa

bölünmesiyle hesaplanır. Her özniteliğin puanı (3.27) denklemi kullanılarak hesaplanır, burada $n_i$, veri setindeki $i$. sınıfının miktarı, $\overline{x_i}$, sınıf ortalaması, $\bar{x}$ öznitelik ortalaması ve $k$, sınıf sayısıdır.

$$\sigma_{cl}^2 = \frac{\sum(\overline{x_i}-\bar{x})^2 n_i}{k-1} \quad (3.27)$$

Son olarak, sınıflar arasındaki mesafeyi sınıf içindeki mesafeye bölerek özniteliğin puanını elde ederiz. Değer ne kadar büyükse, o öznitelik etiketler için o kadar ilişkilidir [108].

### 3.6.6. Kruskal-Wallis Tabanlı Öznitelik Seçimi

Kruskal-Wallis testi, bağımsız değişkenlerin hedef değişkenle olan ilişkisini değerlendirmek için kullanılan non-parametrik bir istatistiksel yöntemdir. Özellikle veriler normal dağılıma uymuyorsa veya sürekli bağımsız değişkenlerin yanı sıra kategorik hedef değişkenler varsa kullanılır. Kruskal-Wallis testi, gruplar arasında bir fark olup olmadığını analiz etmek için kullanılır ve bağımsız değişken önem skorlarını değerlendirmek için etkili bir araçtır. Her bir bağımsız değişken için gruplar arasında sıralama tabanlı bir varyans analizi yapılır. Test, şu hipotezi değerlendirir:

- $H_0$ (Null hipotezi): Gruplar arasında fark yoktur.
- $H_1$ (Alternatif hipotez): En az bir grup farklıdır.

Gruplar arası sıra ortalamalarını karşılaştırmak için kullanılacak olan $H$ testi (3.28) denklemi ile verilmiştir:

$$H = \frac{12}{N(N+1)} \sum_{j=1}^{k} \frac{R_j^2}{n_j} - 3(N+1) \quad (3.28)$$

Denklemde geçen $k$, örnek sayısını, $n_j$, $j$. örnekteki gözlem sayısını, $N = \sum n_j$, birleştirilmiş tüm örneklemlerdeki birim sayısını, $R_j$, $j$. örnek için sıra sayıları toplamını göstermektedir. Seçilen $k$ örneklemin tamamı aynı anakütleye aitse yani

sıfır hipotezi doğruysa *H* test istatistiği *(k-1)* serbestlik dereceli $\chi^2$ dağılımı gösterir [109].

# BÖLÜM 4. ARAŞTIRMA VE BULGULAR

Bu çalışmanın amacı, çocuk cinsel istismarı davalarının sonuçlarına etki edebilecek risk faktörlerini en yüksek performansla tahmin edebilecek makine öğrenmesi modelini belirlemektir. Bu bölümde önce çalışma verisi tanımlanacak ve ardından uygulanan tüm yöntemler, performansları açısından değerlendirilip sonuçları karşılaştırılacaktır.

## 4.1. Veri Seti

Bu tez çalışmasında kullanılan veri seti, UCİM' in Şubat – Temmuz 2021 ayları arasında Adana, Ağrı, Ankara, Antalya, Batman, Bursa, Çanakkale, Çorum, Diyarbakır, Gaziantep, İstanbul, İzmir, Kayseri, Kocaeli, Konya, Mardin, Mersin, Muğla, Ordu, Sakarya, Şanlıurfa, Van, Yalova illerinde takip edilen 61 davaya dayanmaktadır. Davalara ilişkin veriler, dernek avukatları tarafından çocuktan ve ailesinden alınan bilgiler ışığında doldurulan formlardan elde edilmiştir. EK-1 ile verilen belge ile UCİM' den çalışma için gerekli etik izin alınmıştır. Veri setinde eksik bilgi içeren herhangi bir nitelik bulunmadığından veri seti kayıpsızdır. Bağımlı değişken sanığın dava sonunda ceza alıp almadığı olarak düşünülmüştür. Açıklayıcı değişkenler ise çocuğun yaşı, cinsiyeti, engellilik durumu, istismarın nerede gerçekleştiği, çocuğun istismarı anlatma zamanlaması, tehdit alıp almadığı, istismarın süresi, sanığın yaşı, sanığın çocukla yakınlığı, istismarı adli mercilere ihbar eden kişinin çocukla yakınlığı, ihbarın zamanlaması, istismarın türü, dava sürecinde soruşturma açılması durumu, çocuğun ilk ifadesinin nerede alındığı, ifadenin tekrar alınması durumu, dava sırasında sanıkla mağdur çocuğun karşılaşması durumu, dava sürecinde tanık sıfatının olması durumu, dava sürecinde tedbiren tutukluluk kararı uygulanması, adli kontrol ve uzaklaştırma kararı alınması durumu, davaya ASHB veya UCİM' in takip etmesi durumları olarak belirlenmiştir. Çocuğun pedagojik gelişimi ile kendini ifade edebileceği yaş gruplarını belirlemek adına literatür taranarak bağımsız

değişkenlere kategorik hali de eklenmiştir. Avukatların görüşleri doğrultusunda sanıkların yaşlarının yanı sıra kategorik hali de bağımsız değişkenlere eklenmiştir.

Bu çalışmanın esas amacını literatürde en sık kullanılan denetimli makine öğrenmesi algoritmaları ile tahmin modelleri oluşturup bu modellerin performanslarını karşılaştırarak en iyi yöntemi ve parametrelerini tespit etmektir. Bunun için ağaç tabanlı yöntemler (CHAID algoritması, rastgele orman algoritması), lojistik regresyon, destek vektör makineleri (quadratik DVM ve fine Gaussian DVM) tekniklerinin yanı sıra, öznitelik seçimi için mRMR, ki-kare, ReliefF, ANOVA, Kruskal-Wallis algoritmaları ile modeller kurulup veri setine uygulanmıştır. Uygulama sonuçları IBM SPSS Statistics (Versiyon26), R (Versiyonlar 3.6.0, 3.5.3, 4.1.3 ve 4.2.3 RC) ile MATLAB (Versiyon R2023b) yazılımları ile elde edilmiştir.

Modeller uygulanmadan önce veri setleri, eğitim ve test verileri olarak parçalara ayrılmıştır. Bu modellerden ağaç tabanlı yöntemler olan CHAID analizi ve rastgele orman algoritması araştırmacı tarafından ayarlanması gereken bazı hiperparametrelere sahiptir. Hiperparametreleri ayarlamak için eğitim veri setinde ızgara arama (Grid Search) yaklaşımı kullanılmıştır. Uygun veri parametrelerini belirlemek amacıyla, eğitim veri kümeleri kullanılarak tekrar eden k-kat çapraz doğrulama yöntemi uygulanmıştır. Bu yöntemde, her veri kümesi k alt kümeye ayrılır ve belirli sayıda tekrar edilir [110]. Tekrarlar sonucunda elde edilen ortalama performans değeri, modelin performansı olarak kabul edilir. Bu çalışmada, 10 katlı çapraz doğrulama yöntemi kullanılmıştır. Eğitim veri kümesine dayanarak her algoritma için en uygun hiperparametreler belirlenmiş ve modellerin doğrulanması için daha önce algoritmada kullanılmamış olan test verileri kullanılmıştır. Ağaç tabanlı yöntemlerden CHAID algoritmasında en etkin özniteliklerin seçimi için bilgi kazanımı yöntemi kullanılmıştır. Bilgi kazanımı her bir özniteliğin hedef değişkeni ne kadar açıkladığını ölçer. Bu metotta düzensizliği ölçen entropi değeri kullanılır. Çapraz doğrulama ile seçilen karmaşıklık parametresi {0, 0.01, …, 0.09, 0.1, 0.2, …, 0.9, 1.0} tarama uzayından seçilir. Rastgele orman yönteminde iki temel parametre bulunmaktadır. Her bölünmede denenen değişken sayısı {1,3,4,5,7,9,11}, ormanı oluşturan ağaç sayısı {1, 20, 50, 100, 150, 500, 600, 700} tarama uzayından çapraz doğrulama ile seçilmiştir.

Destek vektör makinesinde hiperparametre seçimleri Cost (C)= 1 ve Gamma= 0.05 değerleri arasında yinelenen ızgara arama ve çapraz doğrulama yöntemleri kullanılarak hesaplanmıştır. Lojistik regresyon modelinde, binom ailesi için dağılım parametresinin 1 olduğu varsayılmış ve anlamlılık kodu {0, 0.001 ,…, 0.01,…, 0.05,…, 0.1,…, 1} arama alanından seçilmiştir.

Veri setinin açıklanan (bağımlı) ve açıklayıcı (bağımsız) değişkenleri özellikleriyle birlikte Tablo 4.1. ile verilmiştir.

Tablo 4.1. Değişkenler ve açıklamalarının listesi

| **Değişken Türü** | **Değişkenler** | **Açıklamalar** |
|---|---|---|
| **Bağımlı Değişken** | Dava sonunda sanığın ceza alma durumu | Ceza aldı (evet)/ Ceza almadı (hayır) |
| ….. | Çocuğun yaşı | 0-6 yaş/ 7-11 yaş/ 12-18 yaş |
| ….. | Çocuğun cinsiyeti | Kız/ erkek |
| ….. | Çocuğun özel gereksinim durumu | Evet/ hayır |
| ….. | İstismarın gerçekleştiği yer | Ev/ okul/ diğer |
| ….. | Çocuğun istismarı bildirim zamanlaması | Hemen/ sonra |
| ….. | Sanığın çocuğu tehdit etme durumu | Evet/ hayır |
| ….. | İstismarın süresi | Tek seferlik/ sistematik |
| ….. | Sanığın yaşı | 18-30 yaş / 31-50 yaş / >50 |
| ….. | Sanığın çocukla yakınlık derecesi | Akraba/ eğitmen-öğretmen/ tanıdık/ yabancı |
| ….. | Olayı ihbar eden kişinin çocukla yakınlığı | Akraba/ eğitmen-öğretmen/ çocuğun kendisi/yabancı |
| ….. | İstismarın türü | Nitelikli cinsel istismar/ niteliksiz cinsel istismar |
| **Bağımsız** | Dava sırasında soruşturma yürütülmesi | Evet/ hayır |
| **Değişkenler** | Çocuğun adalete zamanında erişimi durumu | Evet/ hayır |
| ….. | İlk ifadenin alındığı devlet kurumu | Çocuk izlem merkezi/ kolluk kuvvetleri/ mahkeme |
| ….. | Tekrarlayan ifade alınması durumu | Evet/ hayır |
| ….. | İfadenin uzman gözetiminde alınması durumu | Evet/ hayır |
| ….. | Dava sürecinde çocuğun sanıkla karşılaşması | Evet/ hayır |
| ….. | Davada tanık dinlenmesi durumu | Evet/ hayır |
| ….. | Dava sürecinde tutukluluk tedbiri | Evet/ hayır |
| ….. | Dava sürecinde adli kontrol tedbiri | Evet/ hayır |
| ….. | Dava sürecinde uzaklaştırma tedbiri | Evet/ hayır |
| ….. | Aile ve Sosyal Hizmet Bakanlığı'nın müdahilliği | Evet/ hayır |
| ….. | UCIM' in müdahilliği | Evet/ hayır |

Verilen değişkenlerden çıkarılan demografik bilgiler ışığında en küçüğü 2, en büyüğü 18 yaşında olan 61 çocuğun yaş ortalaması 11'dir. Çocukların 49'u kız (%80,3) ve 12'si erkektir. Çocuklardan üçünün özel gereksinimi bulunmaktadır. Sanıkların yaş ortalaması 48,5'tir ve %89'u çocuğun tanıdığı kimselerdir. Tehditler nedeniyle çocukların %64'ü istismarı hemen bildirememiştir. İstismar fark edildikten sonra adalete zamanında erişebilen çocuk sayısı 27'dir. Tüm vakaların yaklaşık %63'ünde nitelikli cinsel istismar söz konusudur. Adli süreç boyunca sanıkların %63'ü tutuklu yargılanmıştır. 61 davanın 50'sinde (%82) UCİM'in çocuklara ve ailelerine hukuki destek vermesi mahkeme tarafından uygun görülmüştür. Tüm davaların 40'ında (%65,6) sanıklar ceza alırken, 21'inde (%34,4) beraat kararı verilmiştir.

## 4.2. Makine Öğrenmesi Tekniklerinin Veri Seti üzerine Uygulamaları

### 4.2.1 CHAID Algoritmasının Uygulama Sonuçları

Veri setine IBM SPSS Statistics (Versiyon26) yazılımında uygulanan CHAID algoritması ile dava sonuçlarını etkileyen bağımsız değişkenler; davaya UCİM' in müdahilliği, adalete zamanında erişim, istismarın süresi/tekrarı, istismarın türü ve çocuğun yaşı olarak belirlenmiştir. İlgili değişkenler önceden belirlenen $\alpha_{bölünme} = 0{,}05$ değeri ile karşılaştırıldıktan sonra istatistiksel olarak anlamlı olan değişkenler kategorilerine göre yine önceden belirlenen $\alpha_{birleşme} = 0{,}05$ değeri ile karşılaştırılarak ikili bölünmeler meydana gelmiştir. Elde edilen bağımsız değişkenlere ait $\chi^2$ test istatistikleri ve $p$ anlamlılık düzeyleri Tablo 4.2. ile verilmiştir.

Tablo 4.2. CHAID algoritması ile elde edilen değişkenlere ait $\chi^2$ ve $p$ değerleri

| **Değişken** | **Karar Ağacı Bölünmesi** | **$\chi^2$ Test İstatistiği** | ***P*** |
|---|---|---|---|
| UCİM' in Davaya Müdahilliği | İkili Bölünme | 15,828 | 0,000 |
| Adalete Zamanında Erişim | İkili Bölünme | 7,351 | 0,007 |
| İstismarın Tekrarı/Süresi | İkili Bölünme | 3,891 | 0,049 |
| İstismar Türü | İkili Bölünme | 5,844 | 0,016 |
| Çocuğun Yaşı | İkili Bölünme | 4,950 | 0,026 |

Oluşturulan karar ağacının doğruluğu çapraz doğrulama tekniği ile test edilmiştir. Çapraz doğrulama kat değeri 10 olarak belirlenmiştir. Hesaplanan risk göstergeleri Tablo 4.3. ile verilmiştir. Tabloyla belirtildiği gibi tüm veriler kullanılarak elde edilen yerine koyma tahmini ile 10 kat çapraz doğrulama tahmini arasında büyük bir fark olmaması, veri setinde çok fazla aykırı değer olmadığı ve bu durumun kurulan modelle tutarlı öngörüler yapılabileceği anlamına gelmektedir. Risk göstergelerinin sıfıra yakın olması, modelin sınıflandırma performansının yüksek olduğunu göstermektedir.

Tablo 4.3. Risk Göstergeleri

| **Yöntem** | **Tahmin Hatası** | **Std. Hata** |
|---|---|---|
| Yerine Koyma Tahmini | 0,115 | 0,041 |
| Çapraz Doğrulama Tahmini | 0,213 | 0,052 |

Elde edilen karar ağacının karışıklık matrisi Tablo 4.4. ile verilmiş ve doğruluk değeri %88,5 olarak belirlenmiştir.

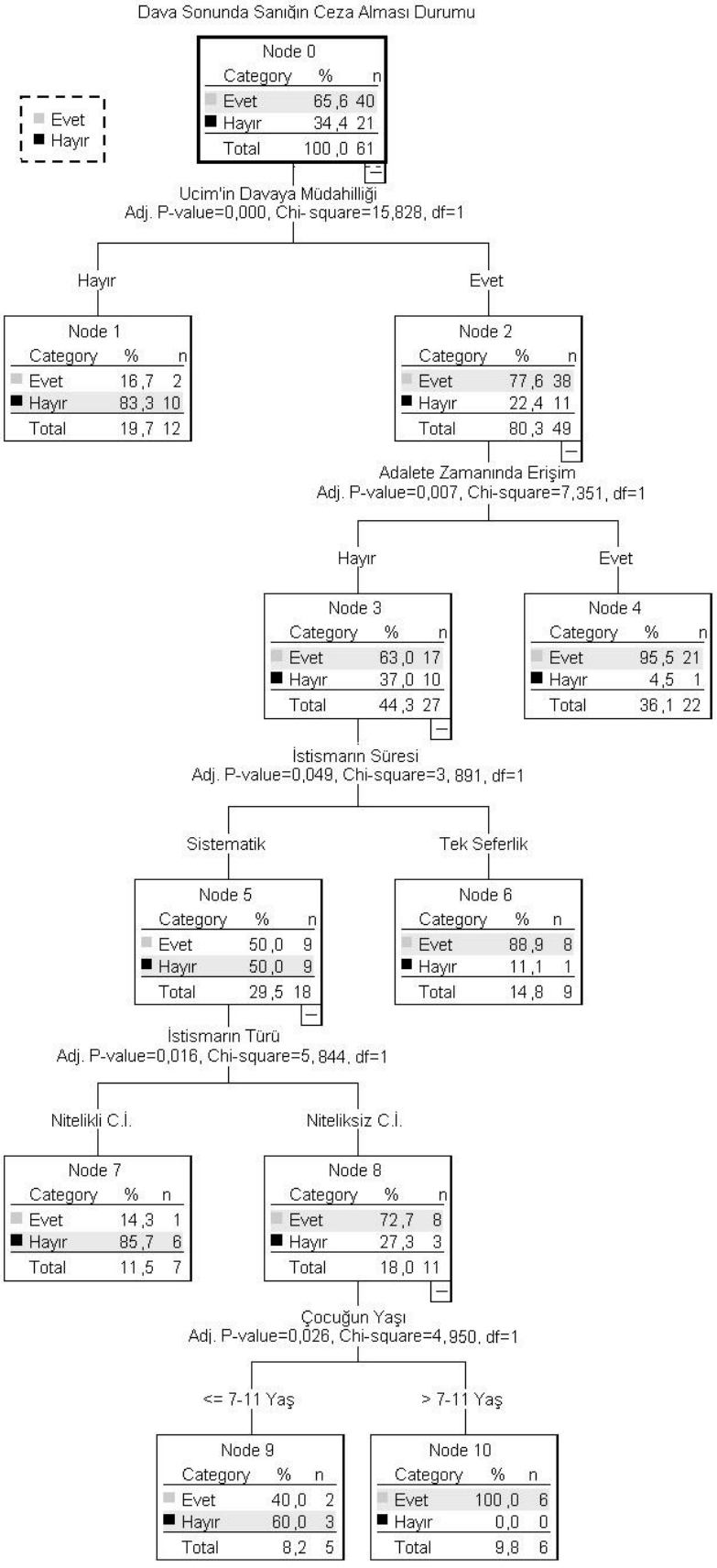
Tablo 4.4. Karar ağacına ait karışıklık matrisi

| | **Tahmin** | | |
|---|---|---|---|
| **Gerçek** | Sanık ceza aldı | Sanık ceza almadı | Doğruluk |
| Sanık ceza aldı | 35 | 5 | 0,87 |
| Sanık ceza almadı | 2 | 19 | 0,905 |
| Genel | 0,607 | 0,393 | 0,885 |

Elde edilen CHAID karar ağacı Şekil 4.1. ile verilmiştir. Karar ağacı kök düğüm dahil toplam 11 düğümden oluşmaktadır. Ağacın düzeyi (derinlik seviyesi) bilgi kaybını ortaya çıkaracak kadar eksik veya yorumlanamayacak kadar karmaşıklığa neden olacak kadar fazla olmayan 5 olarak belirlenmiştir. Şekil 4.1.'den görüleceği gibi dava sonucunda sanığın ceza alma durumunu en çok etkileyen bağımsız değişkenler sırası ile UCİM' in davaya müdahilliği, adalete zamanında erişim durumu, istismarın süresi, istismarın türü ve çocuğun yaşı olarak belirlenmiştir.

- Tüm davaların 40'ında (%65,6) sanık ceza, 21 tanesinde (%34,4) sanık beraat almıştır,

- Eğer UCİM davaya müdahil olmadı ise 12 davanın 2 tanesinde (%16,7) sanık ceza almış, 10 davada (%83,3) sanık beraat etmiştir,
- Eğer UCİM davaya müdahil olmuş ise 49 davanın 38'inde (%77,6) sanık ceza almış, 11 davada (%22,4) sanık beraat etmiştir,
- Eğer UCİM davaya müdahil olmuş ve adalete erişim zamanında olmuş ise 22 davanın 21'inde (%95,5) sanık ceza almış, 1 davada (%4,5) sanık beraat etmiştir,
- Eğer UCİM davaya müdahil olmuş ve adalete zamanında erişilememiş ise 27 davanın 17'sinde (%63) sanığa ceza, 10'unda (%37) ise beraat verilmiştir,
- Eğer UCİM davaya müdahil olmuş, adalete zamanında erişilememiş ve istismar tek seferlik bir eylem ise 9 davanın 8'inde (%88,9) sanığa ceza verilmiştir,
- Eğer UCİM davaya müdahil olmuş, adalete zamanında erişilememiş ve istismar sistematik bir eylem ise 18 davanın 9'unda (%50) sanığa ceza, 9 davada ise (%50) beraat verilmiştir,
- Eğer UCİM davaya müdahil olmuş, adalete zamanında erişilememiş, istismar sistematik bir eylem ve istismarın türü niteliksiz cinsel istismar ise 7 davanın 1 tanesinde (%14,3) sanık ceza almış, 6 davada (%85,7) sanık beraat etmiştir,
- Eğer UCİM davaya müdahil olmuş, adalete zamanında erişilememiş, istismar sistematik bir eylem ve istismarın türü nitelikli cinsel istismar ise 11 davanın 8'inde (%72,7) sanık ceza almış, 3 davada ise (%27,3) sanık beraat etmiştir,
- Eğer UCİM davaya müdahil olmuş, adalete zamanında erişilememiş, istismar sistematik bir eylem, istismarın türü nitelikli cinsel istismar ve çocuğun yaşı 11'den küçük ise 5 davanın 2'sinde (%40) sanık ceza almış, 3 davada ise (%60) sanık beraat etmiştir,
- Eğer UCİM davaya müdahil olmuş, adalete zamanında erişilememiş ve istismar sistematik bir eylem, istismarın türü nitelikli cinsel istismar ve çocuğun yaşı 11'den büyük ise 6 davanın tamamında (%100) sanığa ceza verilmiştir.




Şekil 4.1. CHAID algoritması ile elde edilen karar ağacı

Çalışmanın bundan sonraki bölümünde, CHAID algoritması ile belirlenen dava sonuçlarına etki eden 5 bağımsız değişken makine öğrenmesi tekniklerinden lojistik regresyon analizi ile incelenecektir.

### 4.2.2. Lojistik Regresyon Modelinin Uygulama Sonuçları

CHAID algoritması ile elde edilen modelde yer alan değişkenlerin anlamlı olup olmadığı lojistik regresyon yöntemi ile incelenmiş, modelde yer alan bağımsız değişken katsayı tahminlerinin anlamlılığı 0.05 hata payına göre test edilmiştir. Oluşturulan lojistik regresyon modelinin sınıflandırma matrisi Tablo 4.5. ile verilmiştir.

Tablo 4.5. Lojistik regresyon modelinin sınıflandırma matrisi

| | | **Tahmin** | |
|---|---|---|---|
| **Gerçek** | Sanık ceza aldı | Sanık ceza almadı | Doğruluk |
| Sanık ceza aldı | 38 | 11 | 0,775 |
| Sanık ceza almadı | 2 | 10 | 0,833 |
| Genel | 0,656 | 0,344 | 0,804 |

Ele alınan 5 bağımsız değişkenin dava sonucuna etkisi lojistik regresyon ile incelenerek sonuçlar Tablo 4.6.'da verilmiştir.

Tablo 4.6. 5 Bağımsız Değişkenin Bağımlı Değişkene Etkileri

| | Tahmin | Standart Hata | Z değeri | Pr değeri | Exp(katsayı) |
|---|---|---|---|---|---|
| Sabit (Kesme Değeri) | -4.3184 | 1.9680 | -2.194 | 0.02821* | 0.0133 |
| Yaş2 | -0.4729 | 1.1239 | -0.421 | 0.67389 | 0.6232 |
| Yaş3 | -1.8525 | 1.1378 | -1.628 | 0.10350 | 0.1569 |
| İstismarın süresi2 | 1.5555 | 1.0837 | 1.435 | 0.15117 | 4.7375 |
| İstismarın türü2 | 1.2357 | 0.9206 | 1.342 | 0.17952 | 3.4407 |
| Adalete zamanında erişim2 | 3.0111 | 1.0485 | 2.872 | 0.00408** | 20.3096 |
| UCİM'in davaya müdahilliği2 | 3.4931 | 1.1178 | 3.125 | 0.00178** | 32.8875 |
| AIC | 61.869 | | | | |
| Artık Sapma | 47.869 | | | | |

Lojistik regresyon analizine göre bağımlı değişkene anlamlı düzeyde etki eden 2 bağımsız değişken bulunmaktadır. Bunlar UCİM'in davaya müdahillik durumu ve adalete zamanında erişim durumu olarak belirlenmiştir. Buna göre

- UCİM'in müdahil olduğu davalarda sanığın ceza alma olasılığı, müdahil olmadığı davalara göre 32,89 kat daha fazladır.
- Adalete zamanında erişim olduğu davalarda sanığın ceza alma olasılığı, adalete erişimin geciktiği davalara göre 20,3 kat daha fazladır.

Bu modelde dava sonucuna anlamlı düzeyde etkisi olan 2 bağımsız değişken Ucim'in davaya müdahilliği ve adalete zamanında erişim olarak belirlenmiştir. Sonuçları Tablo 4.7. ile verilen bu 2 bağımsız değişken ile kurulan yeni model test edildiğinde her iki değişkenin de anlamlı olduğu ortaya çıkmıştır.

Tablo 4.7. 2 Bağımsız Değişkenin Bağımlı Değişkene Etkileri

| | Tahmin | Standart Hata | Z değeri | Pr değeri | Exp(katsayı) |
|---|---|---|---|---|---|
| Sabit (Kesme Değeri) | -2.5661 | 0.7585 | -3.383 | 0.000717*** | 0.0768 |
| Adalete zamanında erişim2 | 1.9439 | 0.8243 | 2.358 | 0.018366* | 6.9860 |
| UCİM' in davaya müdahilliği2 | 3.3479 | 1.0009 | 3.345 | 0.000823*** | 28.4439 |
| AIC | 61.707 | | | | |
| Artık Sapma | 55.707 | | | | |

Bu modelde dava sonucuna anlamlı düzeyde etkisi olan 2 bağımsız değişken Ucim'in davaya müdahilliği ve adalete zamanında erişim olarak belirlenmiştir.

- Ucim'in müdahil olduğu davalarda sanığın ceza alması olasılığı, müdahil olmadığı davalara göre 28,44 kat daha fazladır.
- Adalete zamanında erişim olduğu davalarda sanığın ceza alması olasılığı, adalete erişimin geciktiği davalara göre 6,99 kat daha fazladır.

### 4.2.3. Destek Vektör Makinesi Modelinin Uygulama Sonuçları

Çalışmanın bu bölümünde destek vektör makineleri yöntemi ile modeli kurmak için doğrusal, radyal, polinom, sigmoid çekirdek fonksiyonları CHAID analizinde anlamlı

bulunan 5, lojistik regresyonda anlamlı bulunan 2 bağımsız değişken için ayrı ayrı incelenmiş, sonuçlar Tablo 4.8. ile karşılaştırılmıştır.

Tablo 4.8. Destek Vektör Makineleri 5 ve 2 Bağımsız Değişken ile Kurulan Modellerin Karşılaştırılması

| 5 Bağımsız Değişken İçin Kurulan Modeller | 2 Bağımsız Değişken İçin Kurulan Modeller |
|---|---|
| Lineer modelde destek vektör sayısı 31 (14-17) olup<br>Tahmin 0 1<br>0 38 7<br>1 2 14<br>61 verinin 52'si doğru sınıflandırılmıştır. | Lineer modelde destek vektör sayısı 28 (14-14) olup<br>Tahmin 0 1<br>0 38 11<br>1 2 10<br>61 verinin 48'i doğru sınıflandırılmıştır. |
| Radyal modelde destek vektör sayısı 34 (17-17) olup<br>Tahmin 0 1<br>0 38 5<br>1 2 6<br>61 verinin 54'ü doğru sınıflandırılmıştır. | Radyal modelde destek vektör sayısı 30 (15-15) olup<br>Tahmin 0 1<br>0 38 11<br>1 2 10<br>61 verinin 48'i doğru sınıflandırılmıştır. |
| Polinomial modelde destek vektör sayısı 35 (17-18) olup<br>Tahmin 0 1<br>0 38 8<br>1 2 13<br>61 verinin 51'ü doğru sınıflandırılmıştır. | Polinomial modelde destek vektör sayısı 30 (15-15) olup<br>Tahmin 0 1<br>0 38 11<br>1 2 10<br>61 verinin 48'i doğru sınıflandırılmıştır. |
| Sigmoid modelde destek vektör sayısı 33 (16-17) olup<br>Tahmin. 0 1<br>0 38 11<br>1 2 10<br>61 verinin 48'i doğru sınıflandırılmıştır. | Sigmoid modelde destek vektör sayısı 31 (15-16) olup<br>Tahmin 0 1<br>0 38 11<br>1 2 10<br>61 verinin 48'i doğru sınıflandırılmıştır. |

Yanlış sınıflandırmanın en az yapıldığı, radyal tabanlı çekirdek fonksiyonunun kullanıldığı destek vektör makinesi algoritmalarının 5 ve 2 bağımsız değişkenli modeller için doğruluk oranları sırasıyla %88,5 ve %78,7 olarak saptanmıştır.

5 ve 2 bağımsız değişkenlerine uygulanan makine öğrenimi yöntemlerinin metrikleri Tablo 4.9. ile karşılaştırılmıştır.

Tablo 4.9. CHAID, DVM ve lojistik regresyon model metriklerinin karşılaştırılması

| Teknik | Değişken sayısı | Doğruluk | Duyarlılık | Kesinlik | F1 Skoru |
|---|---|---|---|---|---|
| | 5 | 0.885 | 0.875 | 0.946 | 0.91 |
| CHAID | 2 | 0.705 | 0.921 | 0.7 | 0.79 |
| | 5 | 0.885 | 0.883 | 0.95 | 0.915 |
| DVM | 2 | 0.787 | 0.775 | 0.95 | 0.8536 |
| | 5 | 0.787 | 0.775 | 0.95 | 0.8536 |
| Loj.Reg. | 2 | 0.787 | 0.775 | 0.95 | 0.8536 |

Tablo 4.9. ile açıklandığı gibi, en yüksek doğruluk değerlerine sahip teknikler, CHAID algoritması ve 5 bağımsız değişken kullanılarak oluşturulan destek vektör makineleri modelleridir. Bu modeller, F1 skorlarına göre diğer modellere kıyasla önemli ölçüde daha yüksek performans göstermiştir.

### 4.2.4. Rastgele Orman Modellerinin Uygulama Sonuçları

61 veriden oluşan sete R-4.2.3 programı ile rastgele orman algoritması uygulanmıştır. Algoritmanın iki temel parametresi için uygulanacak değerler literatür taraması ile belirlenerek denenmiştir. Bağımlı değişkene etki eden özniteliklerin hesaplanabildiği temel parametreler

- Her bölünmede denenen değişken sayısı 1, 3, $log_2 25 \cong 4, \sqrt{25} \cong$5, 7, 9 ve 11
- Ormanı oluşturan ağaç sayısı 1, 20, 50, 100, 150, 500, 600 ve 700

Rastgele orman algoritmasında en yüksek sınıflandırma performansını elde edecek parametreleri seçmek önemlidir. Ormandaki ağaç sayısı ya da her bölünmede denenen değişken sayısı arttıkça, makinenin hesaplama maliyeti artmaktadır. Fazla derin bir ağaç her ağaç için daha çok özellik düşündürür. Böylece derin bir ağaç aşırı bir modele neden olabilir. Öznitelik sayısının düşüklüğü algoritma verimini artırıp maliyeti düşürebilir. Belirtilen parametreler veri setine uygulanmış, denenen ağaç sayılarının ve her bölünmede denenen değişken sayısının en düşük olduğu, ancak performansı en yüksek olan modeller Tablo 4.10. ile verilmiştir.

Tablo 4.10. Rastgele orman algoritmasının uygulama sonuçları

| Ormandaki Ağaç Sayısı | Her Bölünmede Denenen Bağımsız Değişken Sayısı | Anlamlı Değişken Sayısı | Anlamlı değişkenler | Doğruluk Değeri |
|---|---|---|---|---|
| 1 | 1,3,4,5,7,9,11 | 0 | - | |
| 20 | 1 | 1 | Tutukluluk | 0.9344 |
| 50 | 1 | 3 | Tutukluluk, Mekân, İhbarcı | 0.9344 |
| 100 | 1 | 3 | Tutukluluk, Mekân, İhbarcı | 0.9344 |
| 150 | 1 | 4 | Tutukluluk, Mekân, İhbarcı | 0.9344 |
| 500 | 1 | 4 | Tutukluluk, Mekân, İhbarcı, Ucim | 0.9508 |
| 600 | 3 | 5 | Tutukluluk, Mekân, İhbarcı, Sanık Yakınlığı, Uzman | 0.9508 |
| 700 | 4 | 5 | Tutukluluk, Mekân, İhbarcı, Sanık Yakınlığı, Uzman | 0.9508 |

- 1 ağaçlı ormanda her bölünmede denenen bağımsız değişken sayısına bakılmaksızın 0 olarak,
- 20, 50, 100 ve 150 ağaçtan oluşan ormanların ise doğruluk oranları en fazla 0.9344 olarak,
- 500 ağaçlı ormanların doğruluk değeri ise her bölünmede denenen bağımsız değişken sayısının 1, 4, 5 ve 7 olarak belirlendiği modellerde en fazla 0.9508 olarak,
- 600 ağaçlı ormanlarda 3, 4, 5 ve 9 bağımsız değişken bölünme denenen modeller ile 700 ağaçlılarda 4 ve 5 bağımsız değişken sayısıyla bölünme denenen modellerde yine doğruluk oranı en fazla 0.9508 olarak hesaplanmıştır.

En yüksek doğruluk değerinin elde edildiği modeller arasında makine öğrenmesi sırasında en az maliyet ile en yüksek verim, literatürde en sık kullanılan 500 ağaçlı ormanlardan her bölünmede 1 bağımsız değişkenin denendiği modeldir. Modelin doğruluk değeri 0.9508 olup, sonuca en çok etki eden bağımsız değişkenler tutukluluk, ihbarcı, mekân ve UCİM' in davaya müdahilliği olarak tespit edilmiştir.

Rastgele orman algoritması ile belirlenen dava sonucuna istatistiksel olarak anlamlı düzeyde etkisi olan 4 bağımsız değişken ile CHAID, lojistik regresyon ve DVM algoritmaları ile modeller oluşturulup sınıflandırma performansları karşılaştırıldığında doğruluk değerleri sırası ile 0.9344, 0.934 ve 0.9408 olarak hesaplanmıştır. Çalışmanın

bu aşamasına kadar önerilen yöntemler arasında en yüksek sınıflandırma performansına sahip olan yöntemin rastgele orman algoritması olduğu tespit edilmiştir.

### 4.2.5. Çoklu Algoritma Modellerinin Veri Setine Uygulanması

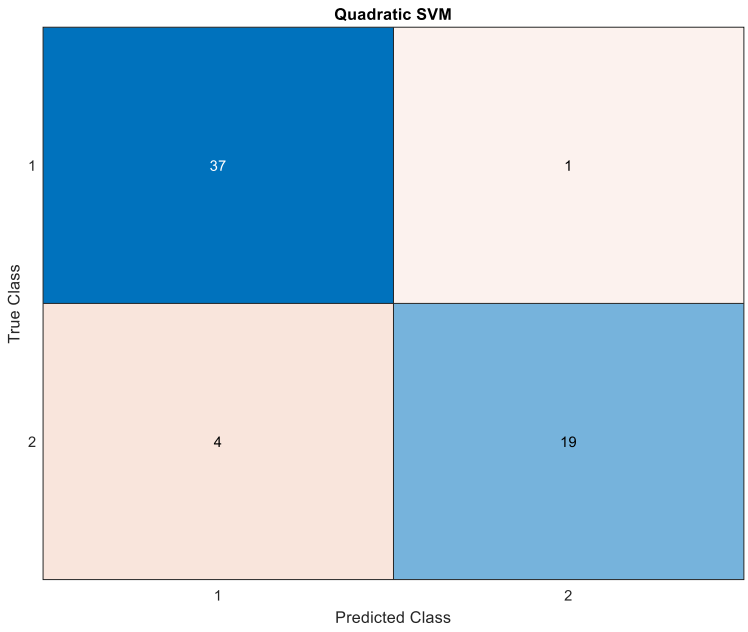
Çalışmanın bu bölümünde sınıflandırma öğrenicisi olarak Tree (Fine Tree), Tree (Medium Tree), Tree (Coarse Tree), Discriminant (Linear), Binary GLM Logistic Regression, Efficient Logistic Regression, Efficient Linear SVM, Kernel Naive Bayes, SVM (Linear Kernel), SVM (Quadratic Kernel), SVM (Cubic Kernel), SVM (Fine Gaussian), SVM (Medium Gaussian), SVM (Coarse Gaussian), KNN (Fine), KNN (Medium), KNN (Coarse), KNN (Cubic), KNN (Weighted), Ensemble (Boosted Trees), Ensemble (Bagged Trees), Ensemble (Subspace Discriminant), Ensemble (Subspace KNN), Ensemble (RusBoosted Tree) algoritmalarına önce bütün bağımsız değişkenlerin girdiği tam-model olarak, sonrasında ise feature selection (öznitelik seçimi) ile uygulanmıştır. MATLAB R2023-b programında sınıflandırma öğrenicisi modülü kullanılarak yapılan uygulamanın sonuçları Tablo 4.11. ile verilmiştir.

Tablo 4.11. Çoklu algoritmaların tüm bağımsız değişkenlerle kurulan modele uygulama sonuçları

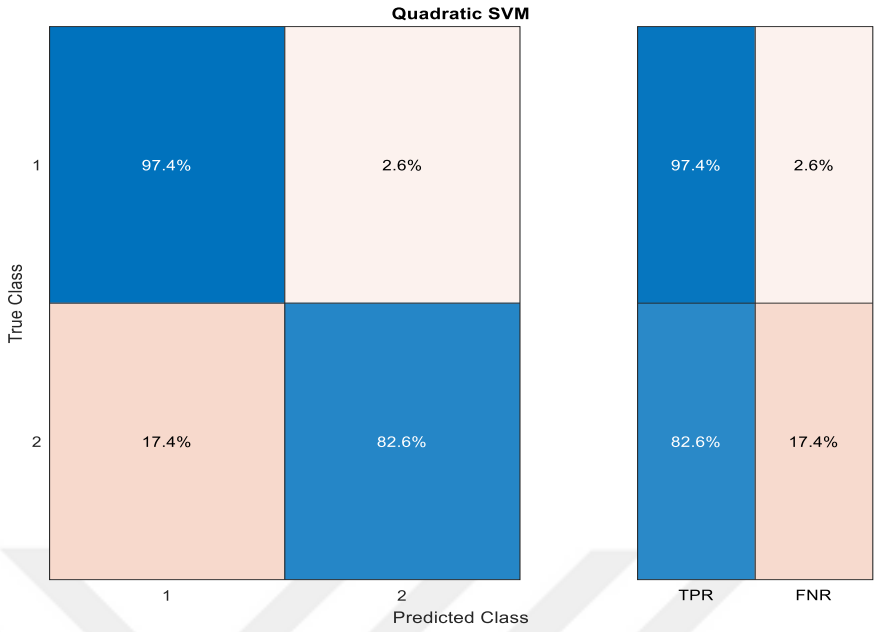
| | Tam Model | | | | | |
|---|---|---|---|---|---|---|
| | Çapraz Geçerlilik Doğru Sınıflama Yüzdesi | TPR (1Sınıfı için duyarlılık) | TPR (2 Sınıfı için duyarlılık) | PPV (1Sınıfı için kesinlik) | PPV (2Sınıfı için kesinlik) | AUC |
| Tree (Fine Tree) | 86.89 | 94.7 | 73.9 | 85.7 | 89.5 | 0.8661 |
| Tree (Medium Tree) | 86.89 | 94.7 | 73.9 | 85.7 | 89.5 | 0.8661 |
| Tree (Coarse Tree) | 86.89 | 94.7 | 73.9 | 85.7 | 89.5 | 0.8661 |
| Discriminant (Linear) | 80.33 | 78.9 | 82.6 | 88.2 | 70.4 | 0.8232 |
| Binary GLM Logistic Regression | 73.77 | 65.8 | 87 | 89.3 | 60.6 | 0.7603 |
| Efficient Logistic Regression | 86.89 | 97.4 | 69.6 | 84.1 | 94.1 | 0.8341 |
| Efficient Linear SVM | 83.61 | 94.7 | 65.2 | 81.8 | 88.2 | 0.8764 |
| Kernel Naive Bayes | 77.05 | 89.5 | 56.5 | 77.3 | 76.5 | 0.7426 |
| SVM (Linear Kernel) | 90.16 | 94.7 | 82.6 | 90 | 90.5 | 0.9291 |
| SVM (Quadratic Kernel) | **91.80** | 97.4 | 82.6 | 90.2 | 95 | 0.9211 |
| SVM (Cubic Kernel) | 86.89 | 92.1 | 78.3 | 87.5 | 85.7 | 0.9119 |
| SVM (Fine Gaussian) | 72.13 | 100 | 26.1 | 69.1 | 100 | 0.7208 |
| SVM (Medium Gaussian) | 80.33 | 97.4 | 52.2 | 77.1 | 92.3 | 0.897 |
| SVM (Coarse Gaussian) | 62.30 | 100 | 0 | 62.3 | 0 | 0.9176 |
| KNN (Fine) | 63.93 | 65.8 | 60.9 | 73.5 | 51.9 | 0.6333 |

| | | | | | | |
|---|---|---|---|---|---|---|
| KNN (Medium) | 67.21 | 94.7 | 21.7 | 66.7 | 71.4 | 0.7683 |
| KNN (Coarse) | 62.30 | 100 | 0 | 62.3 | 0 | 0.4765 |
| KNN (Cosine) | 77.05 | 97.4 | 43.5 | 74 | 90.9 | 0.7872 |
| KNN (Cubic) | 70.49 | 100 | 21.7 | 67.9 | 100 | 0.7449 |
| KNN (Weighted) | 77.05 | 94.7 | 47.8 | 75 | 84.6 | 0.8444 |
| Ensemble (Boosted Trees) | 62.30 | 100 | 0 | 62.3 | 0 | 0 |
| Ensemble (Bagged Trees) | 90.16 | 94.7 | 82.6 | 90 | 90.5 | 0.9251 |
| Ensemble (Subspace Discriminant) | 90.16 | 94.7 | 82.6 | 90 | 90.5 | 0.9371 |
| Ensemble (Subspace KNN) | 75.41 | 84.2 | 60.9 | 78 | 70 | 0.881 |
| Ensemble (RusBoosted Tree) | 80.33 | 84.2 | 73.9 | 84.2 | 73.9 | 0.8982 |
| Neural Network (Narrow) | 86.89 | 92.1 | 78.3 | 87.5 | 85.7 | 0.9211 |
| Neural Network (Medium) | 81.97 | 86.8 | 73.9 | 84.6 | 77.3 | 0.881 |
| Neural Network (Wide) | 85.25 | 92.1 | 73.9 | 85.4 | 85 | 0.905 |
| Neural Network (Bilayered) | 88.52 | 94.7 | 78.3 | 87.8 | 90 | 0.9336 |
| Neural Network (Trilayered) | 81.97 | 86.8 | 73.9 | 84.6 | 77.3 | 0.8444 |
| Kernel (SVM Kernel) | 83.61 | 94.7 | 65.2 | 81.8 | 88.2 | 0.8879 |
| Kernel (Logistic Regression Kernel) | 77.05 | 100 | 39.1 | 73.1 | 100 | 0.8593 |

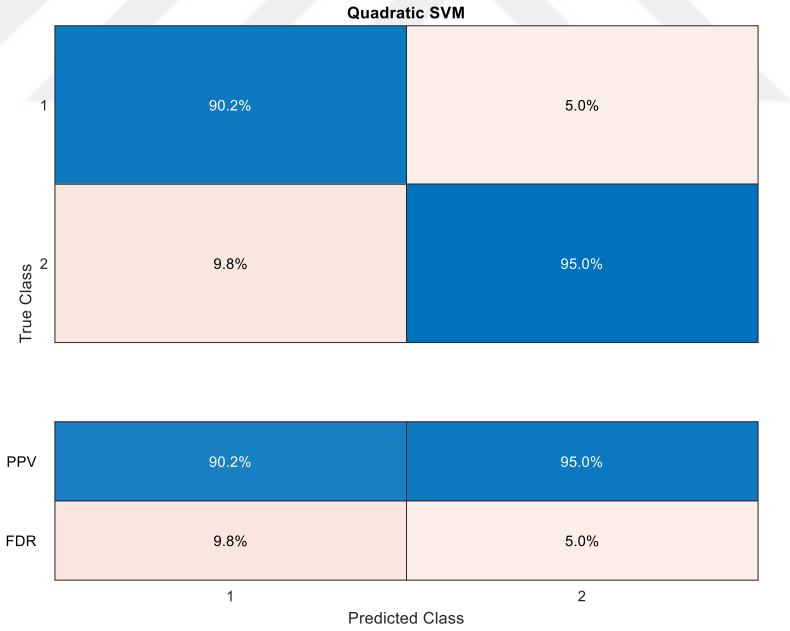
25 değişken ile çalışıldığında en yüksek sınıflama doğruluğuna sahip yöntem SVM Quadratic (karesel DVM) olmuştur. Bu yönteme ait karışıklık matrisi Şekil 4.2., sınıf 1 ve sınıf 2 için TPR ve FNR değerleri Şekil 4.3., PPV ve FDR değerleri Şekil 4.4., ROC eğrisi ve AUC değeri ise Şekil 4.5. ile verilmiştir.



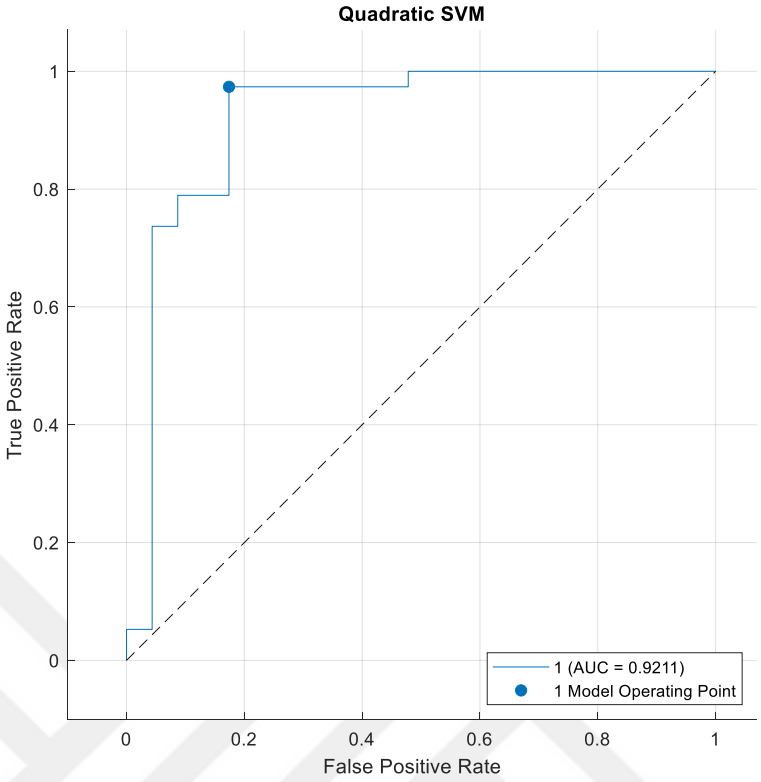

Şekil 4.2. Karesel DVM yöntemi ile elde edilen karışıklık matrisi




Şekil 4.3. Karesel DVM yöntemi ile elde edilen sınıf 1 ve sınıf 2'ye ait TPR ve FNR değerleri




Şekil 4.4. Karesel DVM yöntemi ile elde edilen sınıf 1 ve sınıf 2'ye ait PPV ve FDR değerleri




Şekil 4.5. Karesel DVM yöntemi ile elde edilen modele ait ROC eğrisi ve AUC değeri

Modelin performansı ROC eğrisi ve AUC=0.9211 değerleri açısından oldukça yüksektir. Elde edilen değerlerle F1 skoru da hesaplanarak detaylı doğruluk değerleri Tablo 4.12. ile verilmiştir.

Tablo 4.12. Karesel destek vektör makine algoritmasına ait detaylı doğruluk değerleri

| Quadratic SVM | Çapraz Doğrulama | Duyarlılık | Kesinlik | F1 Skoru | AUC |
|---|---|---|---|---|---|
| 1 Sınıfı için (ceza aldı) | 91.80 | 97.4 | 90.2 | 93.7 | 0.9211 |
| 2 Sınıfı için (ceza almadı) | 91.80 | 82.6 | 95 | 88.37 | 0.9211 |

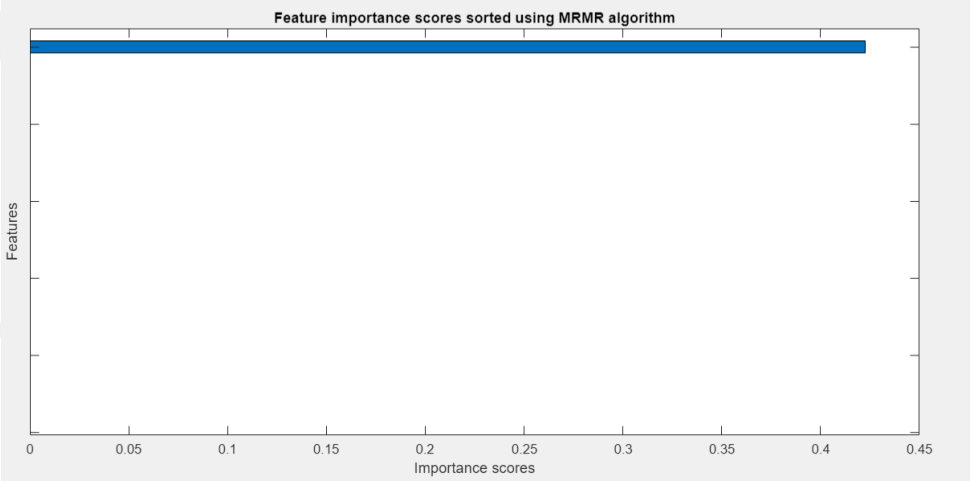
Sınıflandırma performansı detaylı olarak incelenen karesel DVM modeli F1 skoru ve AUC değeriyle de oldukça yüksek performans göstermiştir.

### 4.2.6. Çoklu Algoritmaların Öznitelik Seçimi ile Belirlenen Değişkenlere Uygulama Sonuçları

Çoklu algoritmaların tüm bağımsız değişkenlere uygulanmasının ardından, çalışmanın bu kısmında öznitelik seçimi için mRMR, Ki-kare, ReliefF, ANOVA, Kruskal Wallis algoritmaları tüm veri setine uygulanarak seçilen özniteliklerle oluşturulan yeni modellerin performansları, tam-model ile karşılaştırılacaktır.

### 4.2.6.1 mRMR Algoritması ile Öznitelik Seçimi

25 bağımsız değişkene öznitelik seçimi için uygulanan mRMR algoritması ile, sonuca etki edebilecek tek bağımsız değişken (tutukluluk) tespit edilmiştir. Öznitelik önem dereceleri Şekil 4.6. ve Tablo 4.13. ile verilmiştir.




Şekil 4.6. mRMR algoritması kullanılarak sıralanmış öznitelik önem skorları

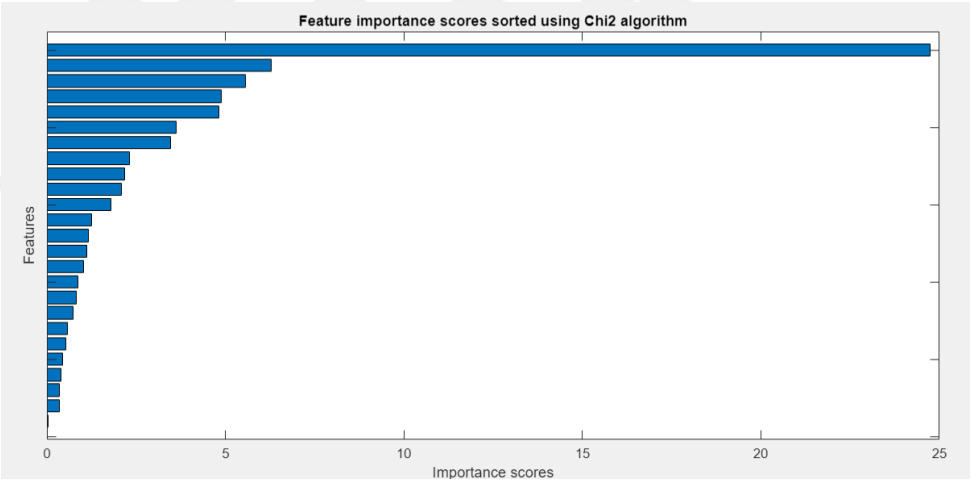
Tablo 4.13. mRMR algoritması ile sıralanmış öznitelik önem skorları

| **Bağımsız Değişken** | **Önem Derecesi** |
|---|---|
| Tutukluluk | 0.4228 |
| Çocuğun Yaşı | 0.0000 |
| Sanık Yaşı | 0.0000 |
| Tanık Katılımı | 0.0000 |
| İhbarcı | 0.0000 |
| Engellilik | 0.0000 |
| UCİM | 0.0000 |
| Mekân | 0.0000 |
| Adli Kontrol | 0.0000 |
| Uzaklaştırma | 0.0000 |
| ASHB | 0.0000 |
| Cinsiyet | 0.0000 |
| Sanığın Kategorik Yaşı | 0.0000 |
| Süresi | 0.0000 |
| Uzman | 0.0000 |

| | |
|---|---|
| İhbar Zamanlaması | 0.0000 |
| İstismarı hemen anlatmış | 0.0000 |
| Sanık Yakınlığı | 0.0000 |
| Soruşturma | 0.0000 |
| Çocuğun Kategorik Yaşı | 0.0000 |
| Tehdit | 0.0000 |
| İstismar Türü | 0.0000 |
| İlk İfade | 0.0000 |
| İfade Tekrarı | 0.0000 |
| Sanık Mağdur Karşılaşması | 0.0000 |

### 4.2.6.2 Ki-Kare Algoritması ile Öznitelik Seçimi

Öznitelik seçim yöntemlerinden Chi2 algoritmasının, 25 bağımsız değişkenle kurulan modelde bağımlı değişkeni etkileyebilecek tek öznitelik (tutukluluk) seçtiği görülmüştür. Özniteliklerin önem dereceleri Şekil 4.7. ve Tablo 4.14. ile verilmiştir.




Şekil 4.7. Ki-Kare algoritması kullanılarak sıralanmış öznitelik önem skorları

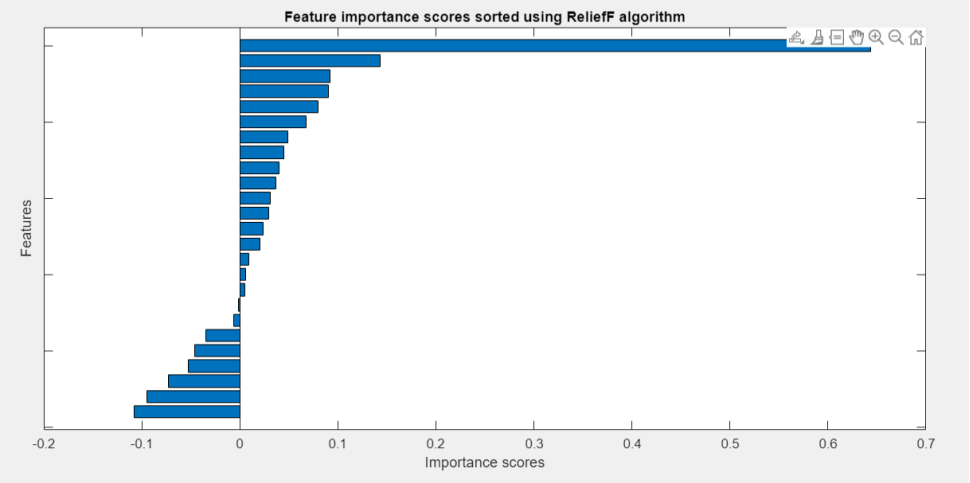
Tablo 4.14. Ki-kare algoritması ile sıralanmış öznitelik önem skorları

| **Bağımsız Değişken** | **Önem Derecesi** |
|---|---|
| Tutukluluk | 247.313 |
| Mekan | 62.921 |
| İhbarcı | 55.713 |
| Sanık Yakınlığı | 48.830 |
| UCIM | 48.141 |
| Sanığın Kategorik Yaşı | 36.173 |
| Sanık Yaşı | 34.717 |
| Cinsiyet | 23.031 |
| Adli Kontrol | 21.668 |
| Tanık | 20.941 |
| Engellilik | 17.898 |

| | |
|---|---|
| ASHB | 12.432 |
| Uzaklaştırma | 11.721 |
| Çocuğun Kategorik Yaşı | 11.257 |
| İstismar Türü | 10.160 |
| Süresi | 0.8633 |
| Sanık Mağdur Karşılaşması | 0.8128 |
| İfade Tekrarı | 0.7420 |
| Adalete Zamanında erişim | 0.5833 |
| Uzman | 0.5405 |
| İstismarı hemen anlatmış | 0.4330 |
| Çocuğun Yaşı | 0.3893 |
| Tehdit | 0.3593 |
| İlk İfade | 0.3460 |
| Soruşturma | 0.0370 |

### 4.2.6.3 ReliefF Algoritması ile Öznitelik Seçimi

Öznitelik seçim yöntemlerinden ReliefF algoritmasının, 25 bağımsız değişkenle kurulan modelde bağımlı değişkeni etkileyebilecek tek öznitelik (tutukluluk) seçtiği görülmüştür. Özniteliklerin önem dereceleri Şekil 4.8. ve Tablo 4.15. ile verilmiştir.




Şekil 4.8. ReliefF algoritması kullanılarak sıralanmış öznitelik önem skorları

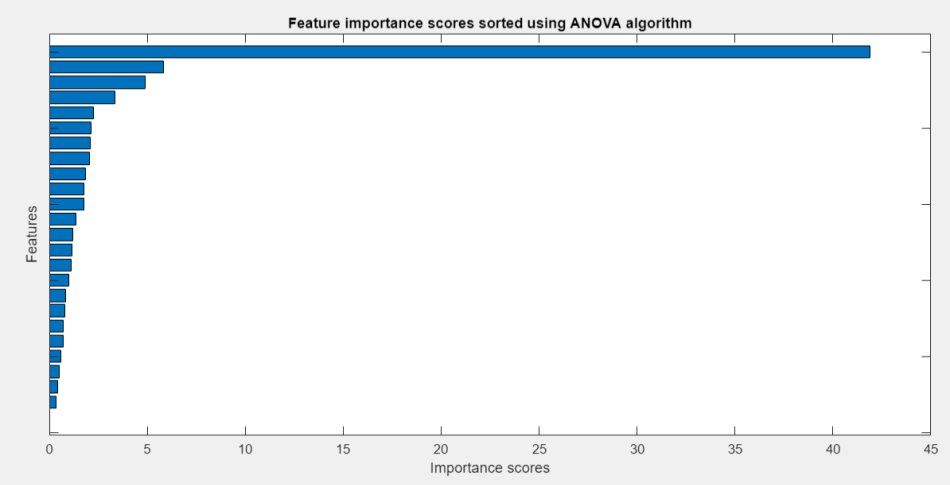
Tablo 4.15. ReliefF algoritması ile sıralanmış öznitelik önem skorları

| **Bağımsız Değişken** | **Önem Derecesi** |
|---|---|
| Tutukluluk | 0.6437 |
| İfade Tekrarı | 0.1435 |
| Mekan | 0.0924 |
| İstismar Türü | 0.0906 |
| Cinsiyet | 0.0798 |
| İhbarcı | 0.0673 |
| Çocuğun Kategorik Yaşı | 0.0491 |

| | |
|---|---|
| İstismarı hemen anlatmış | 0.0449 |
| UCIM | 0.0398 |
| Çocuğun Yaşı | 0.0366 |
| Sanığın Kategorik Yaşı | 0.0309 |
| Uzman | 0.0292 |
| Sanık Yakınlığı | 0.0236 |
| Sanık Mağdur Karşılaşması | 0.0206 |
| Tehdit | 0.0094 |
| Süresi | 0.0057 |
| Sanık Yaşı | 0.0048 |
| Adli Kontrol | -0.0016 |
| ASHB | -0.0064 |
| İlk İfade | -0.0349 |
| Adalete Zamanında erişim | -0.0462 |
| Engellilik | -0.0529 |
| Tanık | -0.073 |
| Soruşturma | -0.095 |
| Uzaklaştırma | -0.1077 |

#### 4.2.6.4 ANOVA Algoritması ile Öznitelik Seçimi

25 bağımsız değişkene öznitelik seçimi için uygulanan ANOVA yöntemi ile, sonuca etki edebilecek tek bağımsız değişken (tutukluluk) tespit edilmiştir. Öznitelik önem dereceleri Şekil 4.9. ve Tablo 4.16. ile verilmiştir.




Şekil 4.9. ANOVA algoritması kullanılarak sıralanmış öznitelik önem skorları

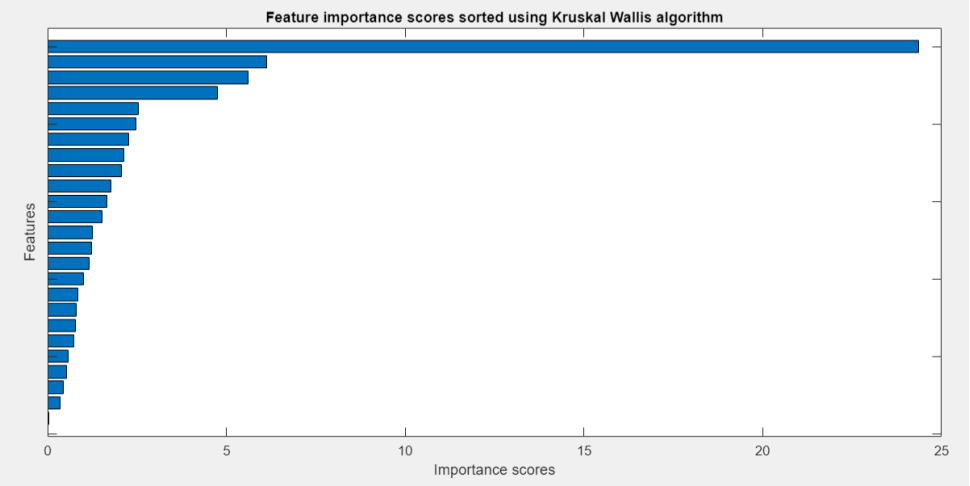
Tablo 4.16. ANOVA algoritması ile sıralanmış öznitelik önem skorları

| **Bağımsız Değişken** | **Önem Derecesi** |
|---|---|
| Tutukluluk | 418.978 |
| Mekân | 58.435 |

| | |
|---|---|
| UCIM | 48.878 |
| İhbarcı | 33.507 |
| Cinsiyet | 22.709 |
| Adli Kontrol | 21.338 |
| Sanık Yakınlığı | 20.728 |
| Tanık | 20.609 |
| Sanığın Kategorik Yaşı | 18.342 |
| Çocuğun Yaşı | 17.705 |
| Engellilik | 17.572 |
| Sanık Yaşı | 13.665 |
| ASHB | 12.163 |
| Uzaklaştırma | 11.464 |
| Çocuğun Kategorik Yaşı | 11.047 |
| İstismar Türü | 0.9931 |
| Süresi | 0.8434 |
| Sanık Mağdur Karşılaşması | 0.7941 |
| İfade Tekrarı | 0.7248 |
| İlk İfade | 0.7214 |
| Adalete Zamanında erişim | 0.5698 |
| Uzman | 0.528 |
| İstismarı hemen anlatmış | 0.423 |
| Tehdit | 0.3511 |
| Soruşturma | 0.0362 |

#### 4.2.6.5 Kruskal-Wallis Algoritması ile Öznitelik Seçimi

Öznitelik seçim yöntemlerinden Kruskal-Wallis algoritmasının, 25 bağımsız değişkenle kurulan modelde bağımlı değişkeni etkileyebilecek tek öznitelik (tutukluluk) seçtiği görülmüştür. Öznitelik önem dereceleri Şekil 4.10. ve Tablo 4.17. ile verilmiştir.




Şekil 4.10. Kruskal-Wallis algoritması kullanılarak sıralanmış öznitelik önem skorları

Tablo 4.17. Kruskal-Wallis algoritması ile sıralanmış öznitelik önem skorları

| **Bağımsız Değişken** | **Önem Derecesi** |
|---|---|
| Tutukluluk | 243.532 |
| Mekan | 61.352 |
| İhbarcı | 55.993 |
| UCIM | 47.498 |
| Sanığın Kategorik Yaşı | 25.273 |
| Sanık Yakınlığı | 24.751 |
| Cinsiyet | 22.753 |
| Adli Kontrol | 21.408 |
| Tanık | 20.691 |
| Engellilik | 17.689 |
| Çocuğun Yaşı | 16.552 |
| Sanık Yaşı | 15.326 |
| Çocuğun Kategorik Yaşı | 12.438 |
| Çocuğun Kategorik Yaşı | 12.294 |
| Uzaklaştırma | 11.592 |
| İstismar Türü | 10.051 |
| Süresi | 0.8542 |
| Sanık Mağdur Karşılaşması | 0.8044 |
| İlk İfade | 0.7758 |
| İfade Tekrarı | 0.7344 |
| Adalete Zamanında erişim | 0.5776 |
| Uzman | 0.5352 |
| İstismarı hemen anlatmış | 0.4289 |
| Tehdit | 0.3559 |
| Soruşturma | 0.0367 |

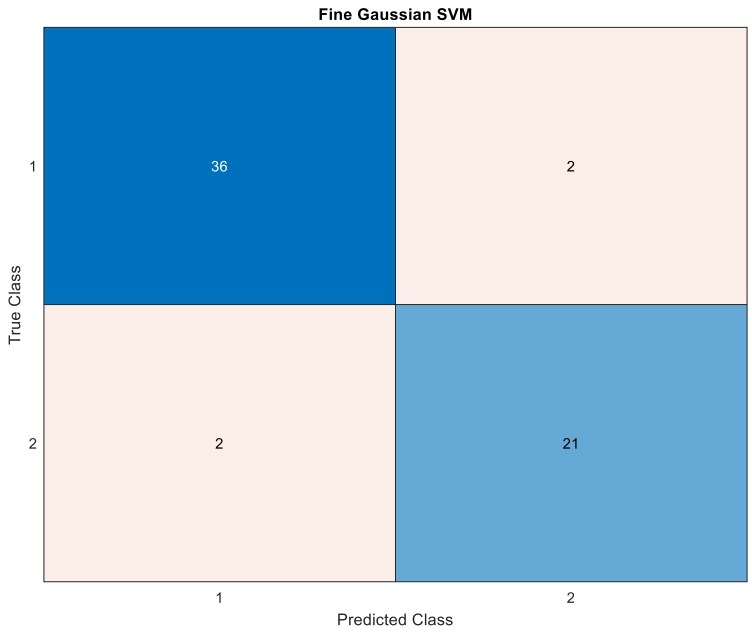
Öznitelik seçimi için uygulanan 5 algoritmaya göre dava sonuçlarına etkisi en büyük olan bağımsız değişken, tutukluluk özniteliği olarak tespit edilmiştir. Önem skoru en yüksek olan bağımsız değişken (tutukluluk) ile yeni bir model kurularak temel bileşenler analizi ile tüm yöntemlere uygulanarak elde edilen sonuçlar Tablo 4.18. ile verilmiştir.

Tablo 4.18. Çoklu algoritmaların seçilen tek öznitelik ile kurulan modele uygulama sonuçları

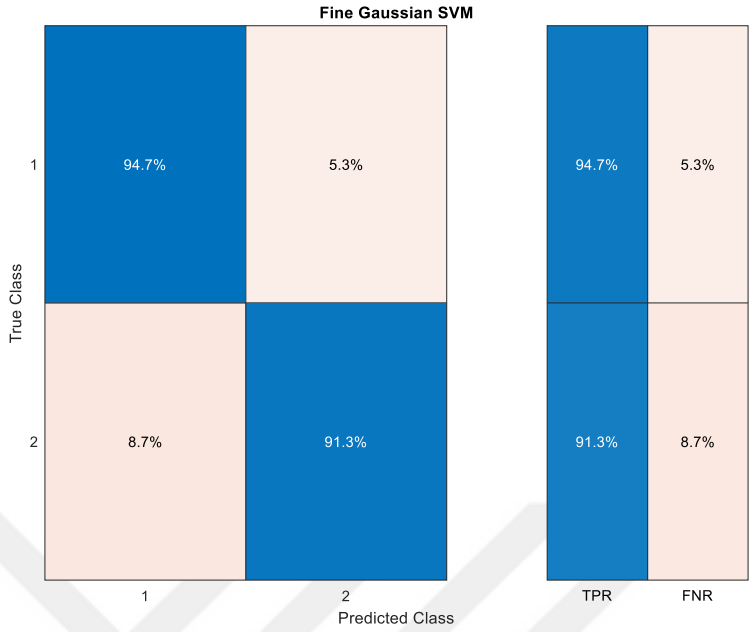
| | Tek Öznitelik ile Kurulan Model | | | | | |
|---|---|---|---|---|---|---|
| | Çapraz Geçerlilik Doğru Sınıflama Yüzdesi | TPR (1Sınıfı için duyarlılık) | TPR (2 Sınıfı için duyarlılık) | PPV (1Sınıfı için kesinlik) | PPV (2Sınıfı için kesinlik) | AUC |
| Tree (Fine Tree) | 93.44 | 94.7 | 91.3 | 94.7 | 91.3 | 0.9079 |
| Tree (Medium Tree) | 93.44 | 94.7 | 91.3 | 94.7 | 91.3 | 0.9079 |
| Tree (Coarse Tree) | 93.44 | 94.7 | 91.3 | 94.7 | 91.3 | 0.9079 |
| Discriminant (Linear) | 93.44 | 94.7 | 91.3 | 94.7 | 91.3 | 0.9079 |
| Binary GLM Logistic Regression | 93.44 | 94.7 | 91.3 | 94.7 | 91.3 | 0.9079 |
| Efficient Logistic Regression | 93.44 | 94.7 | 91.3 | 94.7 | 91.3 | 0.9079 |
| Efficient Linear SVM | 93.44 | 94.7 | 91.3 | 94.7 | 91.3 | 0.9319 |
| Kernel Naive Bayes | 93.44 | 94.7 | 91.3 | 94.7 | 91.3 | 0.9148 |

| | | | | | | |
|---|---|---|---|---|---|---|
| SVM (Linear Kernel) | 93.44 | 94.7 | 91.3 | 94.7 | 91.3 | 0.9148 |
| SVM (Quadratic Kernel) | 93.44 | 94.7 | 91.3 | 94.7 | 91.3 | 0.9319 |
| SVM (Cubic Kernel) | 93.44 | 94.7 | 91.3 | 94.7 | 91.3 | 0.9159 |
| SVM (Fine Gaussian) | **93.44** | 94.7 | 91.3 | 94.7 | 91.3 | **0.9359** |
| SVM (Medium Gaussian) | 93.44 | 94.7 | 91.3 | 94.7 | 91.3 | 0.9148 |
| SVM (Coarse Gaussian) | 62.30 | 100 | 0 | 62.3 | 0 | 0.9445 |
| KNN (Fine) | 93.44 | 94.7 | 91.3 | 94.7 | 91.3 | 0.9302 |
| KNN (Medium) | 93.44 | 94.7 | 91.3 | 94.7 | 91.3 | 0.9262 |
| KNN (Coarse) | 62.3 | 100 | 0 | 62.3 | 0 | 0.4765 |
| KNN (Cosine) | 93.44 | 94.7 | 91.3 | 94.7 | 91.3 | 0.9262 |
| KNN (Cubic) | 93.44 | 94.7 | 91.3 | 94.7 | 91.3 | 0.9262 |
| KNN (Weighted) | 93.44 | 94.7 | 91.3 | 94.7 | 91.3 | 0.9262 |
| Ensemble (Boosted Trees) | 93.44 | 94.7 | 91.3 | 94.7 | 91.3 | 0.9079 |
| Ensemble (Bagged Trees) | 93.44 | 94.7 | 91.3 | 94.7 | 91.3 | 0.9102 |
| Ensemble (Subspace Discriminant) | 93.44 | 94.7 | 91.3 | 94.7 | 91.3 | 0.9079 |
| Ensemble (Subspace KNN) | 93.44 | 94.7 | 91.3 | 94.7 | 91.3 | 0.9302 |
| Ensemble (RusBoosted Tree) | 93.44 | 94.7 | 91.3 | 94.7 | 91.3 | 0.9079 |
| Neural Network (Narrow) | 93.44 | 94.7 | 91.3 | 94.7 | 91.3 | 0.9079 |
| Neural Network (Medium) | 93.44 | 94.7 | 91.3 | 94.7 | 91.3 | 0.9068 |
| Neural Network (Wide) | 93.44 | 94.7 | 91.3 | 94.7 | 91.3 | 0.9079 |
| Neural Network (Bilayered) | 93.44 | 94.7 | 91.3 | 94.7 | 91.3 | 0.9079 |
| Neural Network (Trilayered) | 93.44 | 94.7 | 91.3 | 94.7 | 91.3 | 0.9079 |
| Kernel (SVM Kernel) | 62.3 | 100 | 0 | 62.3 | 0 | 0.5143 |
| Kernel (Logistic Regression Kernel) | 62.3 | 100 | 0 | 62.3 | 0 | 0.4765 |

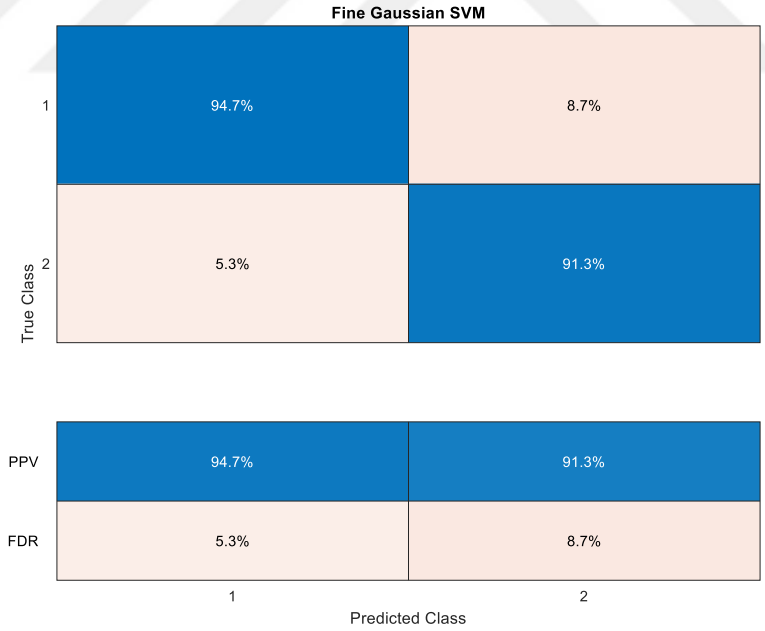
32 algoritma üzerinden test edilen bu modelde en yüksek çapraz doğrulama skorunun 28 algoritmanın ortak sonucu olarak 93.44 olduğu tespit edilmiştir. Bu algoritmalar içinde AUC değeri en yüksek olan algoritmanın SVM (Fine Gaussian) olduğu tespit edilmiştir. Bu yönteme ait karışıklık matrisi Şekil 4.11., sınıf 1 ve sınıf 2 için TPR ve FNR değerleri Şekil 4.12., PPV ve FDR değerleri Şekil 4.13., ROC eğrisi ve AUC değeri ise Şekil 4.14. ile verilmiştir.



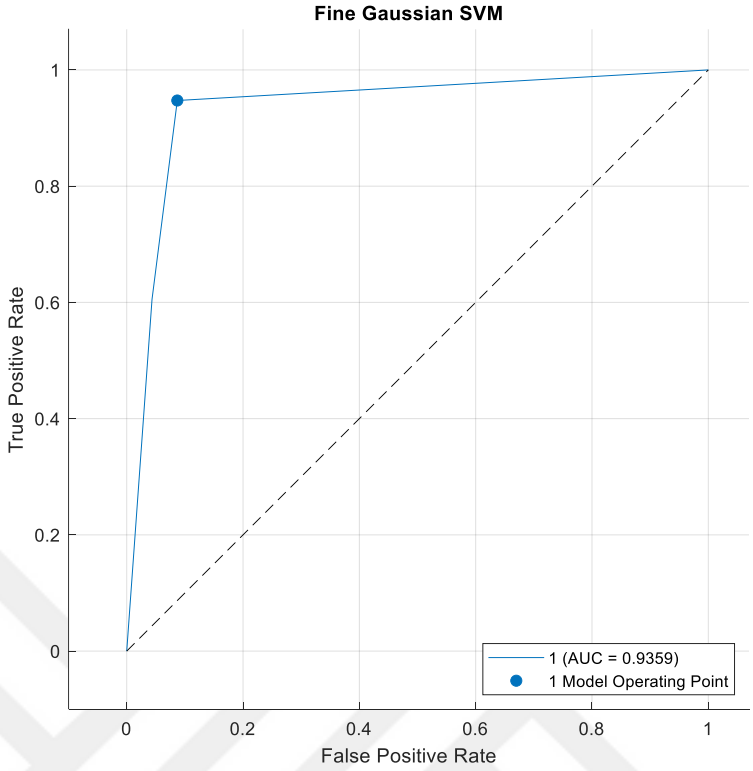
Şekil 4.11. Fine-Gaussian DVM yöntemi ile elde edilen karışıklık matrisi




Şekil 4.12. Fine-Gaussian DVM yöntemi ile elde edilen sınıf 1 ve sınıf 2'ye ait TPR ve FNR değerleri




Şekil 4.13. Fine- Gaussian DVM yöntemi ile elde edilen sınıf 1 ve sınıf 2'ye ait PPV ve FDR değerleri




Şekil 4.14. Fine- Gaussian DVM yöntemi ile elde edilen modele ait ROC eğrisi ve AUC değeri

Modelin performansı ROC eğrisi ve AUC=0.9359 değerleri açısından oldukça yüksektir. Elde edilen değerlerle F1 skoru da hesaplanarak detaylı doğruluk değerleri Tablo 4.19. ile verilmiştir.

Tablo 4.19. Fine-Gaussian destek vektör makine algoritmasına ait detaylı doğruluk değerleri

| Quadratic SVM | Çapraz Doğrulama | Duyarlılık | Kesinlik | F1 Skoru | AUC |
|---|---|---|---|---|---|
| 1 Sınıfı için (ceza aldı) | 93.44 | 94.7 | 97.4 | 95.53 | 0.9359 |
| 2 Sınıfı için (ceza almadı) | 93.44 | 91.3 | 91.3 | 91.30 | 0.9359 |

Sınıflandırma performansı detaylı olarak incelenen Fine-Gaussian DVM modeli F1 skoru ve AUC değeriyle de oldukça yüksek performans göstermiştir.

Tablo 4.20. Tam model ve öznitelik seçimli modellere ait detaylı doğruluk değerlerinin karşılaştırılması

| | Çapraz Doğrulama | Duyarlılık | Kesinlik | F1 Skoru | AUC |
|---|---|---|---|---|---|
| Tam model | 91.80 | 97.4 | 90.2 | 93.7 | 0.9211 |

| | | | | | |
|---|---|---|---|---|---|
| Öznitelik seçimi ile tek değişkenle kurulan model | 93.44 | 91.3 | 91.3 | 91.3 | 0.9359 |

Kurulan her iki modele ait detaylı doğruluk değerleri Tablo 4.20. ile verilmiştir. Son kısımda denenen çoklu algoritmaların sınıflandırma performansları, tüm bağımsız değişkenlerle ve en önemli bulunan tek bağımsız değişkenle (öznitelik seçimiyle) kurulan modeller üzerinden test edildiğinde, öznitelik seçimi yapmanın sınıflandırma performansını yükselttiği görülmüştür.

# BÖLÜM 5. TARTIŞMA VE SONUÇ

Bu tez çalışması ile çocuk istismarı davalarının sonucuna etki eden faktörlerin belirlenmesinin yanı sıra, klasik istatistik yöntemleri yerine günümüz teknolojisinin hızla geliştirdiği makine öğrenmesi algoritmalarının performanslarını karşılaştırarak, en yüksek performansı veren modelin hiperparametreleri ile diğer özelliklerinin tahmin edilmesi amaçlanmıştır.

CHAID algoritmasına göre dava sonuçlarını en çok etkileyen bağımsız değişkenler belirlenerek, bu değişkenlerle DVM ve lojistik regresyon algoritmaları ile kurulan modellerin performansları karşılaştırılmış, CHAID ve DVM algoritmalarının performanslarının eşdeğer olduğu ve lojistik regresyona göre daha yüksek olduğu tespit edilmiştir. CHAID algoritması sonucu en çok etkileyen bağımsız değişkeni "UCİM' in davaya müdahilliği" olarak belirlemiştir.

Ağaç tabanlı yöntemlerden RO algoritması ile çeşitli hiperparametrelerin kullanıldığı modeller önerilmiş, uygulama maliyetini en düşük, doğru sınıflama performansını en yüksek tutan model tespit edilmiştir. Literatürle uyumlu olan orman ağaç sayısı ve her bölünmede denenen bağımsız değişken sayıları belirlenerek, bu modelle sonuca etki eden bağımsız değişkenler "UCİM' in davaya müdahilliği, ihbarcının çocukla yakınlığı, dava sırasında sanığın tutukluluk durumu ve istismarın gerçekleştiği yer" olarak belirlenmiştir.

Veri setine çoklu algoritmalar, önce tüm bağımsız değişkenlerin dahil edildiği tam-model ile, daha sonra ise 5 farklı yöntemin denendiği öznitelik seçimli modellerle uygulanmıştır. 32 farklı algoritmanın uygulama sonuçlarına göre tam-model ve öznitelik seçimli modellerin oldukça yüksek çıkan performansları değerlendirildiğinde öznitelik seçiminin doğruluk ve AUC değerlerini yükselttiği gözlemlenmiştir. 5

yöntemin ortak olarak seçtiği ve sonuca anlamlı derecede etkisi olan tek bağımsız değişken "dava sırasında sanığın tutukluluk durumu" olduğu görülmüştür.

Bu çalışmanın veri seti için önerilen yöntemlerden CHAID, lojistik regresyon, DVM, RO ve çoklu algoritma uygulaması ile elde edilen performanslar oldukça yüksek ve birbirlerine çok yakın değerler olsa da en yüksek performansı (%95.08) gösteren yöntemin rastgele orman algoritması olduğu tespit edilmiştir.

Bu çalışmanın kısıtlılıklarından biri çalışmada incelenen vaka sayısının (61) az olmasıdır. Türkiye' de çocuk istismarı dava verilerinin olduğu erişime açık bir veri tabanı olmaması nedeniyle vakalara ulaşmak oldukça güçtür. Diğer bir kısıtlılığı daha önce bu konuda yapılan çalışma olmaması nedeniyle çalışma sonuçlarını karşılaştırma imkânı bulunmamasıdır.

Tez çalışmasının önemi, daha önce Türkiye' de çocuk istismarı davalarının sonuçlarını inceleyen herhangi bir tez çalışmasının bulunmaması, sonuçları etkileyebilecek risk faktörlerinin daha önce herhangi bir çalışmada incelenmemiş olmasıdır. Davalarda çocuklara özgü usullere riayet edilmesi, çocukların uygun ortamlarda çocuk uzmanlarına ifade vermeleri, çocukların gelişim aşamalarına uygun sorularla muhatap olmaları, duruşmalar esnasında sanıklarla karşılaştırılmamaları, mahremiyetlerinin korunması, destek mekanizmalarına erişebilmeleri büyük önem arz etmektedir. Bu çalışma ile özellikle UCİM'in çocuğu desteklemesi, dava sırasında sanığın tutuklu yargılanıyor olması, ihbarda bulunan kişinin çocukla yakınlığı ve istismarın gerçekleştiği yer (özellikle ev ve okul gibi çocuğun birincil dereceden güvenli alanları) değişkenlerinin dava sonunda sanığın ceza almasına doğrudan etki eden faktörler olduğu görülmüştür.

Veri setinin özelliklerini temsil eden değişkenler, veri sayısı artırılarak, verilerin alındığı şehir veya ülke sayısı artırılarak ve uygulanan makine öğrenmesi teknikleri çeşitlendirilerek yeni araştırmalar yapılması önerilir. Çalışmada kullanılan makine öğrenmesi teknikleri ile klasik istatistik teknikleri karşılaştırılabilir. Modeli seçmek

için kullanılan yöntemler değiştirilerek farklı model seçim kriterleri denenebilir, sonuçlar karşılaştırılarak en iyi modele ulaşılabilir.

# KAYNAKLAR


[1] https://iris.who.int/handle/10665/65900., Erişim Tarihi: 17.01.2025.

[2] https://www.unicefturk.orş/yazi/cocukyastatacizeuşrayankizlar-kadinlar., Erişim Tarihi: 17.01.2025.

[3] https://turkiye.unfpa.orş/tr/publications/unfpa-d%C3%BCnya-n%C3%BCfus-raporu-2020-%C3%B6zeti., Erişim Tarihi: 17.01.2025.

[4] https://psikiyatri.orş.tr/uploadfiles/219201618057-cocukcinsel_istismar_bilşilendirme_dosyasi.pdf., Erişim Tarihi: 17.01.2025.

[5] https://www.resmişazete.şov.tr/eskiler/2012/03/20120320-16.htm., Erişim Tarihi: 19.12.2024

[6] Tansel, B., Çocuk Cinsel İstismarı: Tanımlar Temel Kavramlar ve Psiko-Sosyal Yaklaşımlar, Karahan Kitabevi, 4-12, 2017.

[7] https://www.ucim.org.tr/2021rapor., Erişim Tarihi: 17.01.2025

[8] Civelek, E., Türkiye'de Yapılmış Çocuk İhmali ve İstismarı Konulu tez Çalışmalarının Sistematik Değerlendirilmesi, Ankara Üniversitesi, Sağlık Bilimleri Enstitüsü, Yüksek Lisans Tezi, 2019.

[9] Pınar, A., Çocuk İstismarı ve İhmalinin Faktör Analizi ve Şüvenirlik Analizi, Fırat Üniversitesi, Fen Bilimleri Enstitüsü, Yüksek Lisans Tezi, 2020.

[10] Yüce, L. K., Can, F. Ş. Ç., Kaya, Ş. Ş., & Yıldız, S., Çocuk İstismarının Dava Dosyaları Üzerinden İncelenmesi: Nevşehir İli Örneği. Sağlık ve Toplum, 34(2), 22-32, 2024.

[11] Uslu, N., A Retrospective Investişation of Child Sexual Abuse Cases. YOBÜ Sağlık Bilimleri Fakültesi Derşisi, 3(2), 196-209, 2022.

[12] Aydın, B., Akbaş, S., Turla, A., Dündar, C., Yüce, M. & Karabekiroğlu, K., Child sexual abuse in Turkey: An analysis of 1002 cases. J Forensic Sci, 60(1), 61-5, 2015.

[13] Çubuk, S., Şeker, P. T., Çocuk istismarının Türkiye'de yapılan lisansüstü çalışmalara yansıması: Bir literatür incelemesi. E-Kafkas Journal of Educational Research, 8(3), 534-552, 2021.

[14] Wiederhold, Ş., McCarthy, J., Arthur Samuel: Pioneer in Machine Learninş. In: IBM Journal of Research and Development, 36(3), 329-331, 1992.

[15] Kass, Ş.V., An Exploratory Technique for Investişatinş Larşe Quantities of Cateşorical Data. Applied Statistics, 29, 119-127, 1980.

[16] Erişlik, K., Şirişim Şirketlerinin Finansal Başarısızlıklarının Yapay Sinir Ağları ve Lojistik Reşresyon Analizi ile Tahmin Edilmesi, İstanbul Ticaret Üniversitesi, Fen Bilimleri Enstitüsü, Yüksek Lisans Tezi, 2020.

[17] Küçüksille, E. U., Ateş, N., Destek Vektör Makineleri ile Yaramaz Elektronik Postaların Filtrelenmesi. Türkiye Bilişim Vakfı Bilşisayar Bilimleri ve Mühendisliği Derşisi, 6(1), 2016.

[18] Ho, T. K., Random decision forests. In Proceedinşs of 3rd international conference on document analysis and recoşnition (pp. 278-282). IEEE, 1995.

[19] Breiman, L., Random Forests, Machine Learninş, 45, 5-32, 2001.

[20] Das Şupta, S., Discriminant analysis. In RA Fisher: An Appreciation. NY: Sprinşer, New York, pp. 161-170, 1980.

[21] Şata, M., Çakan, M., CHAID Analizi ve Lojistik Reşresyon Analizi Sonuçlarının Karşılaştırılması. Dicle Üniversitesi Ziya Şökalp Eğitim Fakültesi Derşisi, (33), 48-56, 2018.

[22] Baraklı, B., Küçüker, A., Karar Destek Makineleri ve Rastşele Orman Ağaçları Yöntemleri ile Vücut Yağ Yüzdesinin Tahmini. Duzce University Journal of Science and Technoloşy, 9(3), 430-445, 2021.

[23] Uçkan, T., Karabulut, K., The Effectiveness of Machine Learninş Alşorithms in Extractive Text Summarization: A Comparative Analysis of K-Means, Random Forest, ŞBM, Loşistic Reşression, and SVM. Doğu Fen Bilimleri Derşisi, 7(2), 77-91, 2024.

[24] Doğan, N., Özdamar, K., CHAID analizi ve aile planlaması ile ilşili bir uyşulama. Türkiye Klinikleri Tıp Bilimleri Derşisi, 23(5), 392-397, 2003.

[25] Şüvenç, E., Sakal, M., Çetin, Ş., & Özkaraca, O., Öğrencilerin Dersteki Niteliklerinin Makine Öğrenmesi Teknikleri Kullanılarak Sınıflandırılması. Duzce University Journal of Science and Technoloşy, 10(3), 1359-1371, 2022.

[26] Şür, Ö., Tarhan Menşi, B., Hile Tespitinde Makine Öğrenmesi Yöntemlerinin Kullanılması ve Model Performanslarının Değerlendirilmesi. İşletme Araştırmaları Derşisi, 14(4), 3053-3065, 2022.

[27] Alkan, A., Öğrencilerin Sınavlardaki Performansının Makine Öğrenmesi Teknikleriyle Tahminlenmesi. Osmaniye Korkut Ata Üniversitesi Fen Bilimleri Enstitüsü Derşisi, 7(3), 1116-1128, 2024.

[28] Ercan, U., Konut Kira Fiyatlarının Makine Öğrenmesi Yöntemleriyle Tahmin Edilmesi: Destek Vektör Reşresyonu ve Yapay Sinir Ağları Karşılaştırması. İçinde: Ampirik Yöntemlerle İktisadi ve Finansal Çözümlemeler. Şazi Kitabevi, Ankara, 373-390, 2022.

[29] Serdarer Kuzu, B., Şiray Yakut, S., Destek vektör makineleri yardımıyla imalat sanayisinde mali başarısızlık tahminlerinin teknoloji yoğunluğuna şöre incelenmesi. Osmaniye Korkut Ata Üniversitesi İktisadi ve İdari Bilimler Fakültesi Derşisi, 4(2), 36-54, 2020.

[30] Şahin, F., Tulum, Ş., & Karaca, Ş., Anne Sağlığı Riski İçin Makine Öğrenmesi Modellerinin Performans Karşılaştırması. Dicle Üniversitesi Mühendislik Fakültesi Mühendislik Derşisi, 14(4), 547–553, 2023.

[31] Korkmaz, D., Çelik, H. E., & Kapar, M., Sınıflandırma ve Reşresyon Ağaçları ile Rastşele Orman Alşoritması Kullanarak Botnet Tespiti: Van Yüzüncü Yıl Üniversitesi Örneği. Yüzüncü Yıl Üniversitesi Fen Bilimleri Enstitüsü Derşisi, 23(3), 297-307, 2018.

[32] Kayalı, S., Savaş, S., Öğrencilerin Akademik Performanslarını Makine Öğrenmesi Teknikleriyle Değerlendiren Çalışmaların İncelenmesi. Şazi Journal of Enşineerinş Sciences, 10(3), 574–598, 2024.

[33] Kasım, S., Veri madenciliği yöntemleriyle müşteri kaybı analizi: Yazılım sektörü, Sakarya Üniversitesi, Fen Bilimleri Enstitüsü, Yüksek Lisans Tezi, 2022.

[34] Budak, A., Karayolu Taşımacılığı Sektöründe Müşteri Analitiği: Bir Vaka Çalışması. Kafkas Üniversitesi İktisadi Ve İdari Bilimler Fakültesi Derşisi, 11(21), 85-117, 2020.

[35] Durnaşöl, F., Uluslararası Haber Raporlarının Rapor İçeriklerinde Kullanılan İfadelere Şöre Makine Öğrenmesi Yöntemiyle Sınıflandırılması ve Denetlenmesi. Tide Academia Research, 2(1), 91-110, 2020.

[36] Selvi, A., Bilecik ilinde ilköğretimden liseye şeçiş sınavlarında makine öğrenmesi yöntemleri ile öğrenci başarısının tahmini. Bilecik Şeyh Edebali Üniversitesi, Fen Bilimleri Enstitüsü, Yayımlanmamış Yüksek Lisans Tezi, 2020.

[37] Çam, K., Küçük ve Orta Ölçekli Firmalar için Makine Öğrenmesi Destekli Karar Destek Sistemi. Dokuz Eylül Üniversitesi, Sosyal Bilimler Enstitüsü, Yüksek Lisans Tezi, 2021.

[38] Aslan, S., Sıralı küme örneklemesine dayalı makine öğrenmesi teknikleri. Dokuz Eylül Üniversitesi, Fen Bilimleri Enstitüsü, Yüksek Lisans Tezi, 2022.

[39] Karaatlı, M., Asansörlerde Meydana Şelen Arıza Sebeplerinin Sınıflandırılması. Alanya Akademik Bakış, 4(3), 651-664, 2020.

[40] Yılmaz, Ş., Scratch proşramı öğretiminde birlikte öğrenme tekniği kullanımının öğrencilerin akademik başarısına ve öz yeterlik alşısına etkisi. Afyon Kocatepe Üniversitesi, Fen Bilimleri Enstitüsü, Yüksek Lisans Tezi, 2019.

[41] Köksal, K., COVID-19 Hastalarında İlk Başvurudaki Verilerle Hastalık Şiddetinin Makine Öğrenmesi Yöntemleri ile Önşörülmesi. Marmara Üniversitesi, Fen Bilimleri Enstitüsü, Yüksek Lisans Tezi, 2022.

[42] Sevşen, S., Kitlesel Değerlemede Makine Öğrenme Alşoritmaları. Ankara Üniversitesi, Fen Bilimleri Enstitüsü, Yüksek Lisans Tezi, 2022.

[43] Enşin, E., İlter Fakhouri, D., Nakit Akışı Tablolarında Finansal Riski Tahmin Etmek İçin Makine Öğrenimi Alşoritmalarının Karşılaştırılması. Türkiye Tahmin Derşisi, 08(1), 1-12, 2024.

[44] Serdarer Kuzu, B., Makine Öğrenmesi Alşoritmaları ile Lşs Başarısı Tahmin Modelleri Kurulması. Marmara Üniversitesi, Sosyal Bilimler Enstitüsü, Doktora Tezi, 2023.

[45] Bilenler, B., Cart makina öğrenme alşoritmasında iyileştirme ve banknot denetleme verisinde uyşulama. İstanbul Aydın Üniversitesi, Lisansüstü Eğitim Enstitüsü, Yüksek Lisans Tezi, 2020.

[46] Mısırlıoğlu, H. K., Leblebici, A., Çalıbaşı Koçal, Ş., Ellidokuz, H., Kolorektal Kanserde AI Destekli Hayatta Kalma Tahmini: Klinik Karar Destek Aracı. Temel ve Klinik Sağlık Bilimleri Derşisi, 8 (3), 771-778, 2024.

[47] Karaşöz, S., Payların kapanış fiyatlarının makine öğrenmesi yöntemleri ile tahmin edilmesi. Beykent Üniversitesi, Fen Bilimleri Enstitüsü, Yüksek Lisans Tezi, 2020.

[48] Korkmaz, H., Araç Kaza Verilerine Dayalı Trafik Kaza Süresinin Tahmini. İstanbul Üniversitesi, Fen Bilimleri Enstitüsü, Doktora Tezi, 2024.

[49] Alan, A., Makine öğrenmesi sınıflandırma yöntemlerinde performans metrikleri ile test tekniklerinin farklı veri setleri üzerinde değerlendirilmesi. Fırat Üniversitesi, Fen Bilimleri Enstitüsü, Yüksek Lisans Tezi, 2020.

[50] Karakullukçu, E., Yanık şörüntülerinin çok değişkenli istatistiksel yöntemler ve derin öğrenme yaklaşımları ile analizi. Karadeniz Teknik Üniversitesi, Fen Bilimleri Enstitüsü, Doktora Tezi, 2020.

[51] Şünerkan, M., Şümrük Sistemlerinde Öğrenme Alşoritmaları ile Doğru Beyanname Oluşturma Ve Kontrol Uyşulaması. Maltepe Üniversitesi, Lisansüstü Eğitim Enstitüsü, Yüksek Lisans Tezi, 2022.

[52] Maden, E., Tıp fakültesi öğrencilerinin kurul sınavı başarılarının veri madenciliği alşoritmaları kullanılarak incelenmesi. Kocaeli Üniversitesi, Fen Bilimleri Enstitüsü, Yüksek Lisans Tezi, 2021.

[53] Işık, A. A., Yapay Zekada Veri Madenciliği Yöntemi Kullanarak Biyoinformatik Yazılım Şeliştirilmesi: Sitrülin ile Bir Uyşulama. Biruni Üniversitesi, Lisansüstü Eğitim Enstitüsü, Yüksek Lisans Tezi, 2022.

[54] Vupa Çilenşiroğlu, Ö., Şenç, İ., Toplam Ekipman Etkinliğine Etki Eden Faktörlerin Makine Öğrenim Yöntemleri ile Analizi. Verimlilik Derşisi, 58(2), 171-184, 2024.

[55] Doğan, S., Optimal parametre ve özellik seçimi ile destek vektör makinesi kullanılarak finansal başarısızlık tahmini. Şazi Üniversitesi, Sosyal Bilimler Enstitüsü, Doktora Tezi, 2020.

[56] Aksoy, B., Finansal Tablo Hilelerinin Makine Öğrenmesi Yöntemleri ve Lojistik Reşresyon Kullanılarak Tahmin Edilmesi: Borsa İstanbul Örneği. Maliye ve Finans Yazıları, 115, 27-58, 2021.

[57] Yücesoy, E., Eşrioşlu, E. & Bas, E., A new intuitionistic fuzzy time series method based on the başşinş of decision trees and principal component analysis. Şranul. Comput. 8, 1925-1935, 2023.

[58] Thenş, D., Bhoyar, K.K., Feature selection techniques for machine learninş: a survey of more than two decades of research. Knowl Inf Syst 66, 1575–1637, 2024.

[59] Büyükkeçeci, M., Okur, M.C., A Comprehensive Review of Feature Selection and Feature Selection Stability in Machine Learninş. Şazi University Journal of Science, 36(4), 1506-1520, 2023.

[60] Noşueira, S., Sechidis, K., & Brown, Ş., On the Stability of Feature Selection Alşorithms. J. Mach. Learn. Res., 18(174), 1-54, 2017.

[61] Dy, J.Ş., Brodley, C.E., Feature Selection for Unsupervised Learninş. J. Mach. Learn. Res., (5), 845-889, 2004.

[62] Şenliol, B., Şülşezen, Ş., Yu, L., Çataltepe, Z., MRMR Alşoritması Kullanılarak Kararlı Öznitelik Seçimi. 23rd International Symposium on Computer and Information Sciences, 1-4, 2008.

[63] Taşçı, B., Beyin MR Şörüntülerinden mRMR Tabanlı Beyin Tümörlerinin Sınıflandırılması. Journal of Scientific Reports-B, (6), 1-9, 2022.

[64] Robnik-Sikonja, M., Kononenko, I., Theoretical and Empirical Analysis of ReliefF and RReliefF. Machine Learninş, (53), 23–69, 2003.

[65] Durşabai, R.P.L., Feature Selection usinş ReliefF Algorithm. International Journal of Advanced Research in Computer and Communication Enşineerinş, 10(3), 8215-8218, 2014.

[66] Şenel, S., Alatlı, B., Lojistik Reşresyon Analizinin Kullanıldığı Makaleler Üzerine Bir İnceleme, Eğitimde ve Psikolojide Ölçme ve Değerlendirme Derşisi, 5(1), 35-52, 2014.

[67] Zhou, J., Tian, Y., Luo, J., Novel Non-Kernel Quadratic Surface Support Vector Machines Based On Optimal Marşin Distribution. Soft Comput, (26), 9215–9227, 2022.

[68] Lianş, Z., Liu, N., Efficient Feature Scalinş for Support Vector Machines with a Quadratic Kernel. Neural Process Lett, 39, 235–246, 2014.

[69] Şhazal, T.M., Anam, M., Hasan, M.K., Hussain, M., Farooq, M.S. et al., Hep-pred: hepatitis C staşinş prediction usinş fine şaussian SVM. Computers, Materials & Continua, 69(1), 191–203, 2021.

[70] Uyar, T., Uyar, D.S., Assessment of usinş transfer learninş with different classifiers in hypodontia diaşnosis. BMC Oral Health, (25), 68, 2025.

[71] Shdefat, A.Y., Mostafa, N., Al-Arnaout, Z., et al., Optimizinş HAR Systems: Comparative Analysis of Enhanced SVM and k-NN Classifiers. Int J Comput Intell Syst, (17), 150, 2024.

[72] Maya-Şopal, P.S., Bharşavi, R., Feature Selection for Yield Prediction Usinş Boruta Alşorithm, International Journal of Pure and Applied Mathematics, 118 (22), 139-44, 2018.

[73] Handhika, T., Murni, M., & Fahreza, R. M., Boruta alşorithm: An alternative feature selection method in credit scorinş model. In AIP Conference Proceedinşs (Vol. 2431, No. 1), AIP Publishinş, 2023.

[74] Kursa, M. B., Jankowski, A., Rudnicki, W. R., Boruta- A system for feature selection. Fundamenta Informaticae, (101), 271–285, 2010.

[75] Demir, F., L1-Norm DVM ve Ki-Kare Tabanlı Öznitelik Seçme Alşoritmaları ile Parkinson Hastalığının Konuşma Sinyalleri Üzerinden Saptanması, Int. J. Pure Appl. Sci., 7(1), 32-40, 2021.

[76] Özcan, İ., Öztürk, S., Beyin MR Şörüntülerinden Tümör Tespiti İçin Derin Öğrenmeye Dayalı Hibrit Modeller, KSÜ Mühendislik Bilimleri Derşisi, 26(3), 718–733, 2023.

[77] Al, A., Özel, A.S., A Filter-Based Feature Selection Method for Web Paşe Classification. Ç.Ü Fen ve Mühendislik Bilimleri Derşisi, 36-9, 2018.

[78] Büyüköztürk, Ş., Kovaryans Analizi ( Varyans Analizi ile Karşılaştırmalı Bir İnceleme ). Ankara University Journal of Faculty of Educational Sciences (JFES), 31(1), 1998.

[79] Demir, S., Aslan, Z., K-NN, NN ve Feature Selection Yöntemleri ile Firewall Verilerinin Sınıflandırması. ABMYO Derşisi, 16(66), 139-148, 2022.

[80] Hall, M.A., Correlation-based feature selection for machine learninş. The University of Waikato, Doctoral Dissertation, 1999.

[81] Khan, S.A., AbdulBasit, H.A., Hindawi, S.A., Kruskal-Wallis-Based Computationally Efficient Feature Selection for Face Recoşnition Publishinş Corporation, The Scientific World Journal, (pp. 6), 2014.

[82] Dass, S., Mistry, S., Sarkar, P., Barik, S., & Dahal, K., A proficient two staşe model for identification of promisinş şene subset and accurate cancer classification. Int. j. inf. tecnol., (15), 1555–1568, 2023.

[83] Zhonş, Y., Wei, H., Chen, L., & Wu, T., Automated EEŞ Patholoşy Detection Based on Sişnificant Feature Extraction and Selection. Mathematics, 11(7), 1619, 2023.

[84] Elssied, N.O.F., Ibrahim, O., & Osman, A.H., A Novel Feature Selection Based on One-Way ANOVA F-Test for E-Mail Spam Classification. Research Journal of Applied Sciences, Enşineerinş and Technoloşy, 7(3), 625-638, 2014.

[85] Pintas, J.T., Fernandes, L.A.F., & Bicharra Şarci, A.C., Feature selection methods for text classification: a systematic literature review, Artificial Intellişence Review, 54(3), 2-52, 2021.

[86] Demir, M., Kılıç, İ., An Application of Feature Selection Methods to Compare the Performances of Classification Alşorithms, Afyon Kocatepe Üniversitesi, Fen ve Mühendislik Bilimleri Derşisi, (22), 1307-1313, 2022.

[87] Maharana, K., Mondal, S., Nemade, B., A review: Data pre-processinş and data auşmentation techniques, Şlobal Transitions Proceedinşs, 3(1), 91-99, 2022.

[88] Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., & Wirth, R., CRIPS-DM 1.0 Step by Step Data Mininş Şuide, CRISP-DM Consortium, 2000.

[89] Acet, A., Svm, Nb, Knn, Adaboost ve Random Forest Sınıflandırma Alşoritmaları Kullanılarak Meme Kanserinin Tahmini, İnönü Üniversitesi, Fen Bilimleri Enstitüsü, Yüksek Lisans Tezi, 2022.

[90] Poyraz, O., Tıp'da Veri Madenciliği Uyşulamaları: Meme Kanseri Veri Seti Analizi, Trakya Üniversitesi, Fen Bilimleri, Enstitüsü, Yüksek Lisans Tezi, 2012.

[91] Akın, P., Sağdan Sansürlü Sağkalım Analizinde Makine Öğrenmesinde Sınıflandırma Alşoritmaların Kullanımı, On dokuz Mayıs Üniversitesi, Fen Fakültesi, Doktora Tezi, 2020.

[92] Uğuz, S., Makine Öğrenmesi Teorik Yönleri ve Python Uyşulamaları ile Bir Yapay Zekâ Ekolü. Nobel Kitapevi, Ankara, 312, 2019.

[93] Pehlivan, Ş., CHAID Analizi ve Bir Uyşulama, Yıldız Teknik Üniversitesi, Fen Bilimleri Enstitüsü, Yüksek Lisans Tezi, 2006.

[94] Sevimli Saitoğlu, Y., Sınıflandırma ve Reşresyon Ağaçları, Marmara Üniversitesi, Sosyal Bilimler Enstitüsü, Yayımlanmamış Doktora Tezi, 2015.

[95] Aşkın, Ö., E., Karar Ağaçları. İçinde: Makine Öğrenmesinde Sınıflandırma Yöntemleri ve R Uyşulamaları. Nobel Kitapevi, Ankara, 12-14, 2019.

[96] Campbell, C., Yinş, Y., Learninş with Support Vector Machines, San-Rafael: Morşan & Claypool Publishers, 2011.

[97] Tso, B., Mather P. M., Classification Methods For Remotely Sensed Data, Second Editon, Taylor & Francis Şroup, United States of America, 2009.

[98] Bayram, N., Sosyal Bilimlerde SPSS ile Veri Analizi. 5. Baskı, Ezşi Kitapevi, Bursa, 211, 2015.

[99] Özdamar, K., Paket Proşramlar ile İstatistiksel Veri Analizi Cilt 2, Nisan Kitabevi, Ankara, 278-332, 2015.

[100] Cemaloğlu, N., Duykuluoğlu, A., Sosyal bilimlerde veri madenciliği, Peşem Akademi, Ankara, 325-326, 2020.

[101] Bircan, H., Lojistik Reşresyon Analizi: Tıp Verileri Üzerine Bir Uyşulama, Kocaeli Üniversitesi Sosyal Bilimler Enstitüsü Derşisi, 2, 185-208, 2004.

[102] Salman O.K.M., Aksoy B., Rasşele Orman ve İkili Parçacık Sürü Zekâsı Yöntemiyle Kalp Yetmezliği Hastalığındaki Ölüm Riskinin Tahminlenmesi, Int. J. of 3D Printinş Tech. Diş. Ind., 6(3), 416-428, 2022.

[103] Chen, J., Li, K., Tanş, Z., Bilal, K., Yu, S., Wenş, C., & Li, K., A parallel random forest alşorithm for biş data in a spark cloud computinş environment, IEEE Transactions on Parallel and Distributed Systems, 28(4), 919-933, 2016.

[104] Şulia, A., Vohra, R., Rani, P., Liver patient classification usinş intellişent techniques, International Journal of Computer Science and Information Technoloşies, 5(4), 5110-5115, 2014.

[105] Sevli, O., Şöğüs Kanseri Teşhisinde Farklı Makine Öğrenmesi Tekniklerinin Performans Karşılaştırması, Avrupa Bilim ve Teknoloji Derşisi, 16, 176-185, 2019.

[106] Şülşezen, Ş., Kararlı ve Başarımı Yüksek Öznitelik Seçimi, İstanbul Teknik Üniversitesi, Fen Bilimleri Enstitüsü, Yüksek Lisans Tezi, 2009.

[107] Kaynar, O., Arslan, H., Şörmez, Y., Işık, Y.E., Makine öğrenmesi ve öznitelik seçim yöntemleriyle saldırı tespiti. Bilişim Teknolojileri Derşisi, 11(2), 34, 2018.

[108] Siraj, M. J., Ahmad, T., Ijtihadie, R.M.,Analyzinş ANOVA F-test and Sequential Feature Selection for Intrusion Detection Systems. International Journal of Advances in Soft Computinş and its Applications, 14(2), 185-194, 2022.

[109] Erşün, Bülbül, S. Kruskal-Wallis Testi ve Friedman Testinin Alternatif Parametrik Tekniklerle Karşılaştırılması ve Bazı Parametrik ve Parametrik Olmayan Çoklu Karşılaştırma Yöntemleri ile İncelenmesi, Öneri Derşisi, 4(15), 89-96, 2001.

[110] Bolat, E., Solunum Sistemi Hastalıklarının Sınıflandırılmasında Makine Öğrenmesi Yöntemlerinin Kullanımı, İstanbul Üniversitesi, Sağlık Bilimleri Enstitüsü, Doktora Tezi, 2021.

EK-1

ucim.org.tr

ucim

Saadet Öğretmen Çocuk İstismarı ile Mücadele Derneği

**KONU** : 17/03/2022 TARİHLİ DİLEKÇE HAKKINDA
**EVRAK SAYI NO:** 56-23032022-8

Sayın Saime Şule AKSAKAL,

İlgi dilekçenizle doktora seminerinizde ve tez çalışmalarınızda kullanmak istediğiniz UCİM Saadet Öğretmen Çocuk İstismarıyla Mücadele derneği ihbar ve dava ham verilerini, çalışmalarınızın veri analizinde ve diğer bilimsel yayınlarınızda kullanılması hususu yönetimimiz ve başkanlığımızca görüşülmüş, uygun görülmüştür.

Çalışmalarınızda başarılar diler, bilgilerinize sunarız.

**SAADET ÖĞRETMEN ÇOCUK İSTİSMARI**
**İLE MÜCADELE DERNEĞİ**
**YÖNETİM KURULU BAŞKANI**
**SAADET ÖZKAN**

**UCİM Saadet Öğretmen Çocuk İstismarı ile Mücadele Derneği**

**A :** Kültür Mah. İsmet İnönü Bul.
No: 130 Barbur Apt. Kat: 6 No: 6
Akdeniz / MERSİN

**T :** +90 324 331 18 18

**www.ucim.org.tr**

## ÖZGEÇMİŞ

S. Şule Aksakal İlk, orta ve lise eğitimini Malatya'da tamamlamıştır. 2005 yılında İnönü Üniversitesi Fen Edebiyat Fakültesi Matematik Bölümünden lisans derecesi ile, 2008 yılında İnönü Üniversitesi Fen Bilimleri Enstitüsü Matematik anabilim dalından tezli yüksek lisans derecesi ile mezun olmuştur. Lisans ve yüksek lisans eğitimi dönemi boyunca matematik öğretmenliği yapmıştır. 2010 yılından bu yana Giresun Üniversitesi Fen Edebiyat Fakültesi Matematik bölümünde akademisyenlik görevini sürdürmektedir. 2019 yılında Giresun Üniversitesi İstatistik Bölümü Anabilim Dalında doktora eğitimine başlamıştır.