

**T.C.
MİLLÎ SAVUNMA ÜNİVERSİTESİ
ATATÜRK STRATEJİK ARAŞTIRMALAR VE LİSANSÜSTÜ
EĞİTİM ENSTİTÜSÜ
ENDÜSTRİ MÜHENDİSLİĞİ ANABİLİM DALI
ENDÜSTRİ MÜHENDİSLİĞİ PROGRAMI**

**AŞI TEDARİKİNDE SINIFLANDIRMA
ALGORİTMALARI YARDIMIYLA
GRİP AŞISI TALEBİNİN TAHMİNİ**

YÜKSEK LİSANS TEZİ

**Timuçin ÇINAR
5191102**

TEZ DANIŞMANI: Dr. Öğr. Üyesi Sabahattin Kerem AYTULUN

**İSTANBUL
TEMMUZ 2022**

T.C.
MİLLÎ SAVUNMA ÜNİVERSİTESİ
ATATÜRK STRATEJİK ARAŞTIRMALAR VE LİSANSÜSTÜ
EĞİTİM ENSTİTÜSÜ
ENDÜSTRİ MÜHENDİSLİĞİ ANABİLİM DALI
ENDÜSTRİ MÜHENDİSLİĞİ PROGRAMI

AŞI TEDARİKİNDE SINIFLANDIRMA
ALGORİTMALARI YARDIMIYLA
GRİP AŞISI TALEBİNİN TAHMİNİ

YÜKSEK LİSANS TEZİ

Timuçin ÇINAR
5191102

Tezin Enstitüye Verildiği Tarih:
Tezin Savunulduğu Tarih:

Tez Oy birliği / Oy çokluğu ile başarılı bulunmuştur.

	Unvan Ad Soyad	İmza
Tez Danışmanı	Dr.Öğr.Üyesi Sabahattin Kerem AYTULUN	
Jüri Üyeleri	Dr.Öğr.Üyesi Mehmet YÖRÜKOĞLU	
	Dr.Öğr.Üyesi Hv.Müh.Yb. Ömer ÖZKAN	

İSTANBUL
TEMMUZ 2022

ÖZGÜNLÜK RAPORU

Tez çalışmamın a) Kapak sayfası, b) Giriş, c) Ana bölümler ve ç) Sonuç kısımlarından oluşan toplam 40 sayfalık kısmına ilişkin, 29/06/2021 tarihinde şahsım tarafından Turnitin adlı intihal tespit programından aşağıda belirtilen filtrelemeler uygulanarak alınmış olan özgünlük raporuna göre, tezimin benzerlik oranı %3'dür.

Uygulanan filtrelemeler:

Kaynakça hariç

Alıntılar hariç/dâhil

5 kelimedenden daha az örtüşme içeren metin kısımları hariç

Millî Savunma Üniversitesi Atatürk Stratejik Araştırmalar ve Lisansüstü Eğitim Enstitüsü Lisansüstü Tez Çalışması Özgünlük Raporu Alınması ve Kullanılması Uygulama Usul ve Esasları'nı inceledim ve bu Uygulama Esasları'nda belirtilen azami benzerlik oranlarına göre tez çalışmamın herhangi bir intihal içermediğini; aksinin tespit edileceği muhtemel durumda doğabilecek her türlü hukuki sorumluluğu kabul ettiğimi ve yukarıda vermiş olduğum bilgilerin doğru olduğunu beyan ederim.

Timuçin ÇINAR
29/06/2022

ETİK BEYAN

Millî Savunma Üniversitesi Enstitüleri Lisansüstü Tez Hazırlama Kılavuzu'nda yer alan kurallarına uygun olarak hazırladığım bu tez çalışmasında; tez içinde sunduğum verileri, bilgileri ve dokümanları akademik ve etik kurallar çerçevesinde elde ettiğimi, tüm bilgi, belge, değerlendirme ve sonuçları bilimsel etik ve ahlak kurallarına uygun olarak sunduğumu, tez çalışmasında yararlandığım eserlerin tümüne uygun atıfta bulunarak kaynak gösterdiğimi, kullanılan verilerde herhangi bir değişiklik yapmadığımı, bu tezde sunduğum çalışmanın özgün olduğunu, bildirir; aksi bir durumda aleyhime doğabilecek tüm hak kayıplarını kabullendiğimi beyan ederim.

Bu tezdeki düşünce, görüş, varsayım, sav veya tezler bana aittir; Millî Savunma Bakanlığı, Türk Silahlı Kuvvetleri, Hava Kuvvetleri Komutanlığı, Millî Savunma Üniversitesi ve Atatürk Stratejik Araştırmalar ve Lisansüstü Eğitim Enstitüsü sorumlu tutulamaz.

Timuçin ÇINAR
29/06/2022



Sevgili aileme,

ÖNSÖZ ve TEŞEKKÜR

Öncelikle hayatım boyunca destekleriyle beni hiçbir zaman yalnız bırakmayan sevgili aileme, tez çalışmam süresince bana yol gösteren danışmanım Dr. Öğr. Üyesi Sabahattin Kerem AYTULUN'a ve tezimi yazarken desteğini esirgemeyen arkadaşım Burak Cem KARA'ya sonsuz saygı ve teşekkürlerimi sunarım.

İstanbul; Temmuz 2022

Timuçin ÇINAR



İÇİNDEKİLER

Sayfa

ÖZGÜNLÜK RAPORU	
ETİK BEYANI	
İTHAF	
ÖNSÖZ VE TEŞEKKÜR	
İÇİNDEKİLER	vii
TABLOLAR LİSTESİ.....	vii
ŞEKİLLER LİSTESİ.....	viii
KISALTMALAR	ix
TÜRKÇE ÖZ.....	x
İNGİLİZCE ÖZ (ABSTRACT).....	xi
1. GİRİŞ.....	1
1.1. Tezin Amacı	2
1.2. Tezin Planı	2
1.3. Literatür Araştırması	2
2. VERİ MADENCİLİĞİ VE SINIFLANDIRMA PROBLEMİ.....	5
2.1. Veri Madenciliği	5
2.2. Veri Madenciliği İşlevleri	8
2.3. Sınıflandırma Problemi	9
2.4. Tek Etiketli Sınıflandırma.....	10
2.4.1. İkili Sınıflandırma.....	10
2.4.2. Çok Sınıflı Sınıflandırma.....	10
2.5. Çok Etiketli Sınıflandırma	11
2.5.1. Algoritma Uyarlama Yöntemi	12
2.5.2. Problem Dönüştürme Yöntemi	12
2.6. Sınıflandırma Algoritmaları	15
2.6.1. k En Yakın Komşuluk Algoritması (KNN).....	15
2.6.2. Karar Ağacı Algoritması (DET)	16
2.6.3. Naive Bayes Algoritması (NAB)	17
2.6.4. Lojistik Regresyon Algoritması (LOG)	18
2.6.5. Genelleştirilmiş Doğrusal Model Algoritması (GLM).....	18
2.6.6. Derin Öğrenme Algoritması (DEL)	18
2.6.7. Rassal Orman Algoritması (RAF).....	18
2.6.8. Gradyan Güçlendirilmiş Ağaçlar Algoritması (GBT).....	19
3. BAŞARI DEĞERLENDİRME ÖLÇÜTLERİ.....	20
3.1. Karmaşıklık Matrisi (Confusion Matrix)	20
3.2. Doğruluk (Accuracy).....	21

	Sayfa
3.3. Kesinlik (Precision).....	21
3.4. Duyarlılık (Recall).....	21
3.5. F Puanı (F Score).....	22
3.6. Eğri Altında Kalan Alan (Area Under the Curve - AUC).....	22
4. UYGULAMA.....	23
4.1. Veri Seti Özellikleri	23
4.2. Veri Ön İşleme	23
4.3. İkili Uygunluk Yöntemi Uygulaması.....	25
4.4. Etiket Kuvvet Kümesi Yöntemi Uygulaması.....	28
4.5. Sınıflandırıcı Zincirler Yöntemi Uygulaması	28
4.6. Sonuç ve Bulgular	29
5. SONUÇ VE ÖNERİLER.....	30
KAYNAKÇA	31
EKLER.....	33
Ek 1. Veri Seti Özellikleri.....	34
Ek 2. BR Tahminlerinin Karmaşıklık Matrisleri	38
Ek 3. CC Tahminlerinin Karmaşıklık Matrisleri	39
ÖZGEÇMİŞ.....	40

TABLÖLAR LİSTESİ

	Sayfa
Tablo 3.1: Karmaşıklık matrisi	20
Tablo 4.1: Çok eksik değer içeren öznitelikler	23
Tablo 4.2: Nominal değer içeren öznitelikler	24
Tablo 4.3: H1N1 grip aşısı olma tahmini performansı	25
Tablo 4.4: Mevsimsel grip aşısı olma tahmini performansı	25
Tablo 4.5: Etiket kuvvet kümesi yöntemi tahmin performansı	26
Tablo 4.6: Sınıflandırıcı zincirler yöntemi h1n1 grip aşısı olma tahmini performansı.....	29
Tablo 4.7: Sınıflandırıcı zincirler yöntemi mevsimsel grip aşısı olma tahmini performansı	29
Tablo 4.8: Karşılaştırma tablosu.....	29

ŞEKİLLER LİSTESİ

	Sayfa
Şekil 2.1: Veri madenciliği kullanım alanları.....	5
Şekil 2.2: Veri madenciliği süreci	6
Şekil 2.3: Veri ambarı kullanım çatısı	7
Şekil 2.4: Bir veri küpü örneği	8
Şekil 2.5: Sınıflandırma çeşitleri ve yöntemleri	9
Şekil 2.6: İkili sınıflandırma örneği.....	10
Şekil 2.7: Çok sınıflı sınıflandırma örneği	11
Şekil 2.8: Çok etiketli veri seti örneği	11
Şekil 2.9: İkili uygunluk yöntemi örneği.....	13
Şekil 2.10: Etiket kuvvet kümesi yöntemi örneği.....	14
Şekil 2.11: Sınıflandırıcı zincirler yöntemi örneği	14
Şekil 2.12: k en yakın komşu algoritması örneği	15
Şekil 2.13: Karar ağacı yapısı örneği.....	16
Şekil 2.14: Rassal orman yapısı örneği.....	19
Şekil 3.1: ROC AUC grafikleri	22
Şekil 4.1: RapidMiner Studio veri ön işleme süreç akışı.....	24
Şekil 4.2: Impute Missing Values operatörü süreci.....	24
Şekil 4.3: RapidMiner Studio ikili uygunluk yöntemi süreç akışı.....	25
Şekil 4.4: Data process operatörü süreci.....	26
Şekil 4.5: Karar ağacı algoritması operatörü süreci.....	26
Şekil 4.6: Karar ağacı algoritması “Optimize Parameters” operatörü süreci	26
Şekil 4.7: Karar ağacı algoritması “Cross Validation” operatörü süreci	27

KISALTMALAR

AUC	: ROC Eğrisi Altındaki Alan
BR	: İkili Uygunluk
CC	: Sınıflandırıcı Zincirler
COVID-19	: Koronavirüs Hastalığı
DEL	: Derin Öğrencme Algoritması
DET	: Karar Ağacı Algoritması
GBT	: Gradyan Güçlendirilmiş Ağaçlar Algoritması
GLM	: Genelleştirilmiş Doğrusal Model Algoritması
H1N1	: Domuz Gribi
KNN	: k En Yakın Komşuluk Algoritması
LOG	: Lojistik Regresyon Algoritması
LP	: Güç Kuvvet Kümesi
NAB	: Naive Bayes Algoritması
RAF	: Rassal Orman Algoritması
ROC	: Alıcı İşletim Karakteristiği
VM	: Veri Madenciliği

ÖZ

Aşı Tedarikinde Sınıflandırma Algoritmaları Yardımıyla

Grip Aşısı Talebinin Tahmini

Timuçin ÇINAR

Millî Savunma Üniversitesi, Atatürk Stratejik Araştırmalar ve

Lisansüstü Eğitim Enstitüsü

İstanbul, (Temmuz, 2022)

Aşılar, hastalıklarla mücadele için önem arz etmektedir. Çünkü aşılama tüm canlılar için bağışıklık sağlar. Bunun yanında sürü bağışıklığı yöntemi ile de salgınların yayılımı azaltılabilir. Aşı tedariki yapılırken tahminlerin iyi yapılamaması, hastalıklarla mücadelede zorluklara sebebiyet vermektedir. Bu çalışmada katılımcılara ekonomik ve demografik bilgileri, hastalık riskleri ve aşıların etkinlikleri hakkındaki görüşleri ve hastalık bulaştırmayı azaltmaya yönelik davranışları sorularak anket bilgileri toplanmıştır. Çalışmada 2009 yılında gerçekleşen H1N1 grip aşısının kamuya açılması sonrasında kişilere yapılan anket verileri ile talep tahminlerinin gerçeğe yakın bir şekilde planlanmasına rehberlik edilmesi amaçlanmıştır. Ayrıca çalışmada makine öğrenme tekniklerinden faydalanarak bireylerin aşı olup olmadığına yönelik tahminler yapılmıştır.

Anahtar Sözcükler: Çok Etiketli Sınıflandırma, İkili Uygunluk, Güç Kuvvet Kümesi, Sınıflandırıcı Zincirler.

Bilim Kodu: 90618

Sayfa Sayısı: 40

Tez Danışmanı: Dr. Öğr. Üyesi Sabahattin Kerem AYTULUN

ABSTRACT

Estimation of Flu Vaccine Demand with
the Help of Classification Algorithms in Vaccine Supply

Timuçin ÇINAR

National Defense University, Atatürk Strategic Studies And
Graduate Institute
Istanbul, (July, 2022)

Vaccines are important to combat diseases. Because vaccination provides immunity for all living things. In addition, the spread of pandemics can be reduced by the herd immunity method. The inability to make good predictions while procuring vaccines causes difficulties in the fight against diseases. In this paper, questionnaire information was collected by asking the participants about their economic and demographic information, their views on disease risks and the effectiveness of vaccines, and their behaviors to reduce disease transmission. In the paper, it is aimed to guide the planning of demand forecasts close to reality with the survey data made to the people after the H1N1 flu vaccine was made public in 2009. In addition, in the paper, predictions were made about whether individuals were vaccinated by using machine learning techniques.

Keywords: Multilabel Classification, Binary Relevance, Label Powerset, Classifier Chains.

Science Code: 90618

Pages: 40

Supervisor: Asst. Prof. Dr. Sabahattin Kerem AYTULUN

1. GİRİŞ

Aşılar, insanlar veya hayvanlar için hastalıklara karşı bireysel bağışıklık sağlar. Bir topluluğun yeterli miktarı bağışıklık sağlandığında, sürü bağışıklığı yoluyla hastalıkların yayılması daha da azaltılabilir.

2009 yılının baharında başlayan ve halk arasında domuz gribi (H1N1) olarak adlandırılan virüsünün neden olduğu bir salgın dünyayı kasıp kavurmuştur. Araştırmacılar ilk yılda dünya çapında 151.000 ila 575.000 insanın bu salgından dolayı öldüğünü tahmin etmektedir (DRIVENDATA, 2021).

Bu tip salgınları önlemek için kullanılan teknolojilerden en önemlisi aşıdır. Aşının salgının yayılmasını engellemesi amacıyla öncelikle toplumun bir plan dahilinde aşılması ve aşının gerektiği yerde gerektiği miktarda hazır bulundurulması sağlanmalıdır. Bu amaçla aşının tedarik edilmesi ve ihtiyaç duyulan yere herhangi bir bozulmaya uğramadan ulaştırılması, söz konusu aşıya olan talebin doğru bir şekilde tespit edilmesine bağlıdır. 2019 yılında başlayan koronavirüs hastalığı (COVID-19) salgını ile mücadele sürecinde haber başlıklarında aşılardan farklı sebeplerle kullanılamaz hale geldiğine de şahit olunmuştur. Bu sebeple akademik literatürde özellikle aşının taşınması sürecinde yaşanan sıkıntılara değinilmiştir (Çevik, 2021). Bu bir sorun olmakla birlikte beraberce çözümlenmesi gereken diğer bir sorun da talebin gerçekçi bir şekilde hesaplanmasıdır.

Günümüzde talep tahmini dendiğinde ilk akla gelen yaklaşım zaman serilerinin çözümlenmesidir. Sabit, trend ve/veya mevsimsellik içeren zaman serilerinin çözümlenmesinde istatistiksel yaklaşımlar kullanılır. Ancak zaman serileri ile bir sonuca varabilmek için serinin geçmiş değerlerine ihtiyaç duyulur. Koronavirüs gibi daha önce yaşanmamış salgınlarda geçmiş bilgilere ulaşmak mümkün olmadığı gibi aynı zamanda enfeksiyonların karakteristiklerindeki farklılıklar sebebiyle de uygun değildir. Bu sebeple zaman serilerinden ziyade insan davranışlarına dayalı analizlerin yapılmasının daha doğru olacağı kanattına ulaşılmıştır. Yapılan anketlerin sonuçları ve/veya devletin yetkili kurumlarında saklanan kişisel bilgi ve tercihler kullanılarak

veri madenciliği yoluyla aşı talebinin daha gerçekçi olarak tespiti mümkün olabilecektir.

Veri madenciliği günümüzde bankacılık, telekomünikasyon, sağlık, bilimsel araştırmalarda sıklıkla başvurulmuş bir disiplin haline gelmiştir. Özellikle veri madenciliğinin önemli bir alt konusu olan sınıflandırma yöntemleri bir tahmin sürecinde önemli katkılar sağlamaktadır. Örneğin; bir müşterinin sadık bir müşteri olup olmadığı, dolandırıcılık tespiti, hastalık teşhisi gibi konular günümüzde bilim insanlarının çokça çalıştığı alanlardır.

1.1 Tezin Amacı

Pandemik hastalıklarda etkili bir aşının en kısa zamanda geliştirilmesi ve toplumun yeterli kısmına aşılanmanın yapılması büyük bir önem arz etmektedir. Ancak aşı tedariki yapılırken aşı taleplerinin doğru tahmin edilememesinden ötürü yüzbinlerce aşı çöpe gitmektedir.

Bu çalışmada, aşılanma öncesi yapılacak anket çalışmaları ile elde edilen ve/veya devletin yetkili kurumlarında var olan bireysel bilgilerin veri madenciliği kullanılarak işlenmesi sayesinde, aşılanma bölgelerine gönderilecek olan aşı tedarik miktarlarının gerçeğe uygun tahmin edilerek kullanılamayan aşı miktarının en aza indirilmesi ve toplumun aşılanma miktarının maksimize edilmesi amaçlanmaktadır.

1.2 Tezin Planı

İkinci bölümde veri madenciliğine genel bir bakıştan sonra sınıflandırma problemlerinin ele nasıl alındığından bahsedilmiştir. Üçüncü bölümde değerlendirme ölçütlerinin tanımlamaları ve açıklamalarına yer verilmiştir. Dördüncü bölümde yapılan uygulamalar ve başarı sonuçları gösterilmiştir.

Son olarak sonuç bölümünde tezin genel sonucu özetlenmiş ve konuyla ilgili gelecek çalışmalarda yapılabilecek geliştirmelere yer verilmiştir.

1.3 Literatür Araştırması

Read vd. (2009) çalışmalarında ikili uygunluk (binary relavence - BR) yönteminin etiketler arası ilişkileri göz ardı etmesi, hali hazırda bulunan yöntemlerin etiketler arası ilişkileri kullanarak değerlendirme yapmak için yüksek karmaşıklığa sahip modeller

oluşturması sebeplerinden ötürü büyük veri setleriyle ölçeklenebilir olan sınıflandırıcı zincirler (classifier chains - CC) yöntemini sunmuşlardır. Türlü veri setleri ve ölçütler kullanarak gerçekleştirdikleri deneysel çalışmalarda yöntemin mevcut yöntemler ile rekabet edebileceğini göstermişlerdir.

Tsoumakas vd. (2009) çalışmalarında BR yönteminin doğrusal karmaşıklığa sahip olduğunu ancak etiketlerin arasındaki ilişkileri hesaplamaya dahil etmediğinden ötürü etiket kombinasyonlarının doğru olarak tahmin edilemeyebileceğinden bahsedilmiş ve bu sorunun istifleme (stacking) yöntemiyle kolaylaştırılabileceğine değinmişlerdir. Çalışma dahilinde istifleme yönteminin makine öğreniminde kullanımına dair bilgiler verilmiş ve çok etiketli sınıflandırma için 2BR diye adlandırdıkları bir istifleme yöntemini önermişlerdir. Buna ek olarak çalışmalarında istifleme yönteminin ilk katmanı (base-level) için budama işlemi yapılmasını öne sürmüşlerdir. Bu sayede öznitelik boyutu azalırken sistem etkinliğinin ve performansının arttığını açıklamışlardır.

Montañes vd. (2014) çalışmalarında bağımlı ikili uygunluk yöntemini ortaya koymuşlardır. Bu yöntemin hem zincirleme hem de istifleme yöntemlerinin özelliklerini içerdiğini açıklamışlardır. Çalışma kapsamında önerilen yöntem ile mevcut yöntemler deneysel olarak çok etiketli veri setleri ve çok etiketli sınıflandırma yaygın olarak kullanılan metrikler ile değerlendirilmiş ve bahsi geçen yöntemin diğerlerine göre daha fazla başarımlı gösterdiğini kanıtlamışlardır.

Chen vd. (2021) çalışmalarında bağımlı ikili uygunluk yönteminin başarımlı yükseltmek için istifleme yöntemine ek olarak etiket seçimini de gerçekleştiren Etiket Seçimli İstifleme (Stacking Model with Label Selection) adlı yöntemi ortaya koymuşlardır. Yöntem ilk katmanda BR yöntemi ile aynı mantıkla çalışmakta lakin ikinci katmanda her bir etiket için bir etiket seçim işlemi yaparak her bir etiket için en uygun etiketlerin kümesini ayırarak ikinci katmanın öznitelik kümesini bu etiketler ile birlikte genişletmektedir. Çalışma kapsamında önerilen yöntem deneysel olarak diğer yöntemlerle karşılaştırılmış ve yöntemin mevcut yöntemlerle rekabet edebileceği değerlendirme sonucu kanıtlanmıştır. Çalışmada ortaya koyulan yöntemin etkisinin ve gücünün yüksek olduğu belirtilmiştir.

Cherman vd. (2011) çalışmalarında BR yönteminin bir parçası olacak şekilde BR+ isimli bir yöntem oluşturmuşlardır. Bu yöntemin iki çeşidi mevcuttur. İlki BR+ No Update, DBR yöntemiyle çok benzerlik göstermektedir. Yöntem içerisinde iki adet sınıflandırıcı katman bulunmakta, ilk katman BR yöntemiyle aynı, ikinci katmanda ise

her bir etiket için diğer tüm etiketler kullanılarak özellik uzayını genişleten bir model oluşturulmuştur. İkinci varyasyon BR+ Update Order ise bu yöntem ile zincirleme yaklaşımını birleştirerek ikinci katmandaki her bir modelden elde edilen tahminler ile bir sonraki model içerisindeki etiket değeri güncellenerek başarıyı arttırmak amaçlanmıştır. Çalışma kapsamında ortaya koyulan yöntem ile iki diğer çok etiketli sınıflandırma yöntemi karşılaştırılmış ve ortaya koyulan yöntemin diğer iki yönteme göre iyileştirme sağladığı açıklanmıştır.

Li ve Zhang (2014) çalışmalarında kontrollü etiket ilişkilerinden yararlanmayı gerçekleştirmek için, iki aşamalı bir filtreleme prosedürüne dayalı olarak BR için yeni bir geliştirme önerilmiştir. Önerilen yöntemin ilk aşamasında veri setinin geçiş için ayrılmış bir parçası ve F puanı (F measure) metriği kullanılarak bir budama işlemi gerçekleştirilir. İkinci adımda ise budanmamış etiketler özellik olarak kullanılır ve öznelik seçimi yöntemi ile yakın ilişkili etiketler belirlenir. Çalışma kapsamında on dört çok etiketli veri setinde kapsamlı deneyler gerçekleştirilmiş, özellikle etiket alanında çok sayıda sınıf etiketi bulunduğunda, kontrollü etiket ilişkilerinden yararlanmanın üstünlük sağladığı açıklanmıştır.

Tenenboim vd. (2010) çalışmalarında çok etiketli veri seti içindeki ilişkileri tespit eden ve bu ilişkileri modelleyen bir yaklaşım önerilmiştir. Çalışmalarında ilk olarak etiketler arası ilişkileri tespit eden, devamında etiket kümesini birbirini dışlayan birkaç alt kümeye bölen ve elde edilen etiketler arası ilişkiler ile çok etiketli sınıflandırma işlemi gerçekleştiren bir algoritma geliştirmişlerdir. Çalışmada bağımsız etiketleri tespit etmek için ki-kare testi kullanılmıştır. Keşfedilen ilişkileri içeren çok etiketli sınıflandırma için ikili uygunluk ve etiket güç seti yöntemlerinin bir kombinasyonu kullanılmıştır.

2. VERİ MADENCİLİĞİ VE SINIFLANDIRMA PROBLEMİ

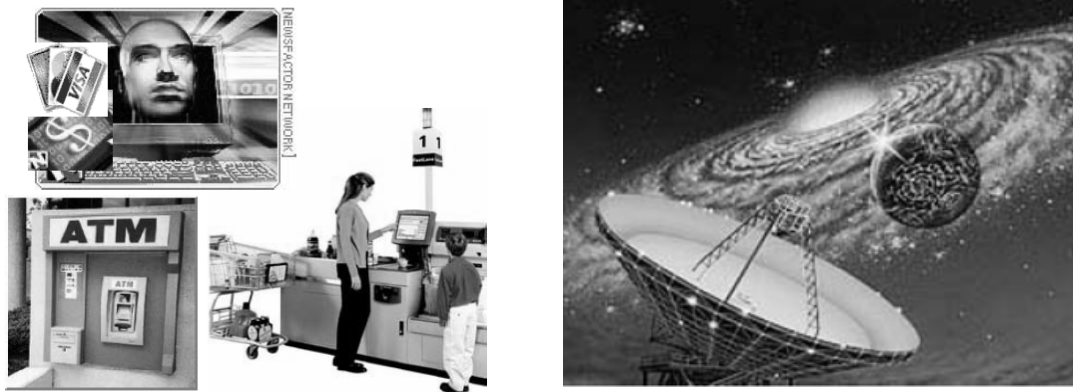
2.1 Veri Madenciliği

Aslında veri madenciliği (VM) konusuna girmeden önce bilgi kavramı üzerinde durmak gerekir. Bilgi temel olarak işaretlerle başlayıp daha sonra ardı arkasına veri, malumat, bilgi olarak devam eden, eyleme dönüştüğünde gücüne kavuşan bir olgudur. Bu bağlamda bugünün şartlarında bilgi teknolojileri ve buna bağlı olarak gelişen yazılımlar ile özellikle bilginin gücü hissetmedilmektedir. Bu bileşenler birlikte bir değer yaratma sürecini tanımlamaktadır (Tan vd., 2014).

Bu değer ekleme sürecinde ilgi alanımızdaki bilgi bazen açık bazen kapalı olarak karşımıza çıkmaktadır. Temel problem kapalı bilginin açık bilgi formuna dönüştürülmesidir. Bunun için bilim dallarının hemen hepsinden faydalanılmaktadır. Örneğin bir bitki, bakteri veya virüsün tanımlanmasında ilgili bütün akademik bilgi yanında matematik ve istatistik disiplinleri de devreye girmektedir.

Çok büyük miktarlarda verinin elde olması ve bu verinin kullanışlı ve anlamlı bilgiye dönüştürülmesi ihtiyacı, VM'ni çok çekici bir hale sokmuştur. Bu anlamlı bilgi, pazar araştırmalarından müşteri tutmaya, üretim kontrolünden bilimsel gelişmelere kadar bir çok alanda çok kullanışlı olmaktadır. Bu sebeple VM, bilgi teknolojilerinin doğal evriminin bir sonucu olarak görülmektedir (Öztemel, 2012).

Bu durum hem ticari açıdan hem de bilimsel açıdan önemli avantajlar sağlamaktadır.

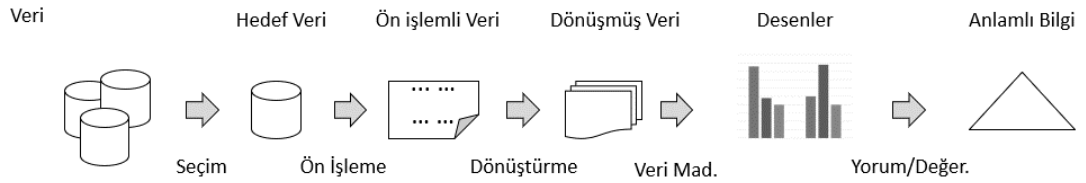


Şekil 2.1: Veri madenciliği kullanım alanları.

Ticari açıdan WEB, e-ticaret, satınalma, banka/kredi kartları, telekomünikasyon vb. yollarla elde edilen bilgiler müşteri ilişkileri yönetimi, satış ve pazarlama stratejilerinin belirlenmesi gibi konularda kullanılmakta iken uydularda bulunan uzaktan algılama sensörleri teleskopik gökyüzü taramaları, gen kodlarının çözümü, terabaytlarca bilgi üreten bilimsel simülasyonlar ile bilimsel açıdan önemli kazanımlar sağlanmaktadır (Tan vd., 2014).

Veri bolluğu, güçlü veri analiz araçları ile çok daha büyük boyutlara ulaşmış ve sonuçta verinin bol ancak bilginin yetersiz olduğu bir durum ortaya çıkmıştır. Çok büyük kapasiteli bilgi depolarında bulunan büyük miktarda ve hızla artma eğilimindeki verinin analizi artık çok güçlü araçlar olmadan yapılamaz hale gelmiştir. Dolayısıyla büyük miktardaki veriler, bir veri mezarlığı oluşturmuş, veri arşivleri çok nadir ziyaret edilen yerler haline almıştır. İşte tam da bu noktada veri madenciliği yöntemleri geliştirilerek bu imkansız görünen veriden bilgiye dönüşümü yapılabilir hale getirmiştir.

Veri madenciliği “Önemsiz olmayan, daha önceden bilinmeyen ve potansiyel olarak kullanışlı bilginin veriden çıkarılması.” veya “Anlamlı desenleri keşfetmek için büyük miktarda verinin otomatik veya yarı otomatik yollarla araştırılması ve analizi.” şeklinde tanımlanabilir (Cerrito, 2006). Bu anlamda Şekil 2.2’de veri madenciliği süreci gösterilmektedir.

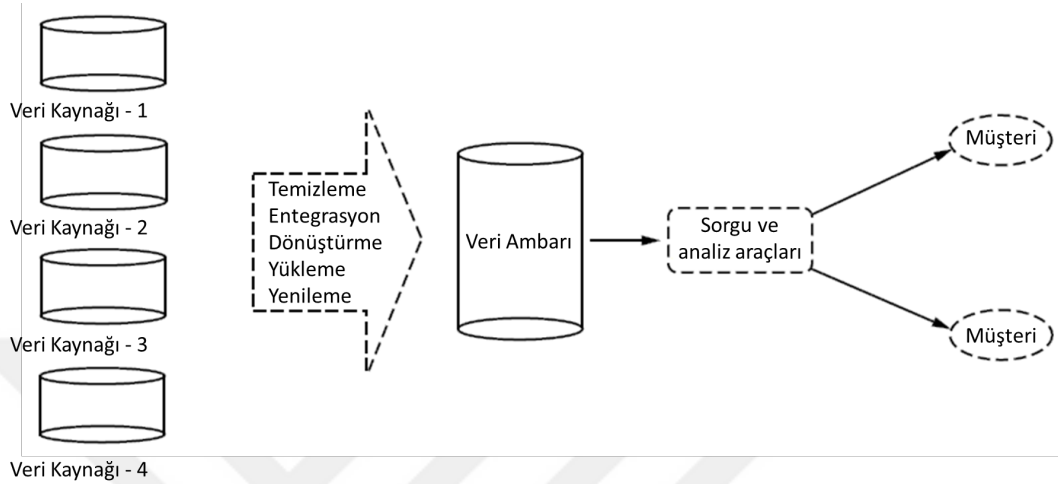


Şekil 2.2: Veri madenciliği süreci.

Öncelikle analize konu olan verinin seçimi yapılmalı, verideki korelasyon, eksiklik, hata vb. sorunlardan arındırılmalı, gerekiyorsa birleştirme ve dönüştürme yapılmalıdır. Bütün bu işlemlere veri ön işleme olarak bilinmektedir. Bu aşama belki de bu sürecin en önemli adımıdır çünkü bu adım yeteri kadar iyi yapılmaz ise yapılan analizler sonucunda elde edilen bilginin geçerliliği de şüpheli olmaktadır. Bu veriye veri

madenciliği algoritmaları uygulandıktan sonra elde edilen sonuçların anlaşılır bir şekilde açıklanması ve uygulamaya verilmesi gerekmektedir.

Veri madenciliği çalışmaları işlemsel veri tabanlarında yapılmaz. Bunun yerine veri ambarları kullanılmaktadır. Veri ambarı kullanım çatısı Şekil 2.3’de verilmiştir.



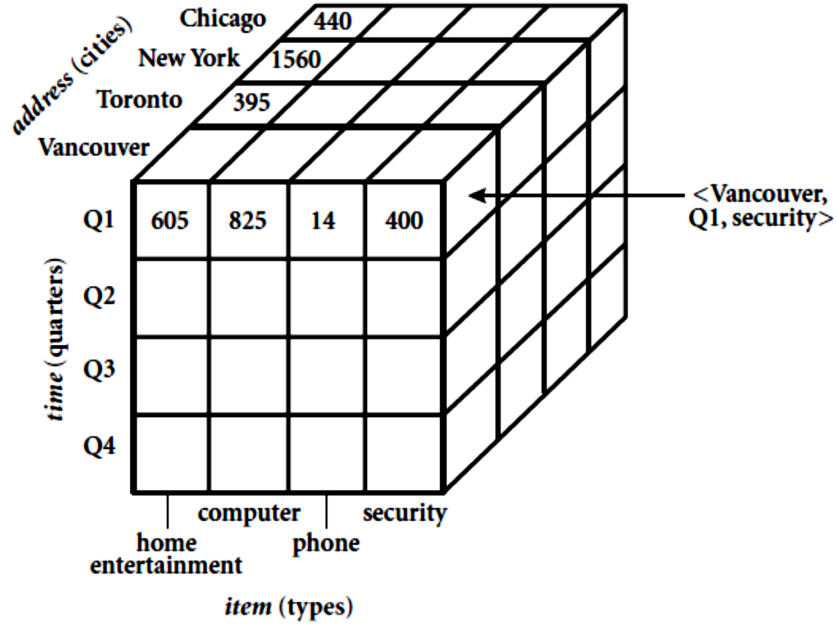
Şekil 2.3: Veri ambarı kullanım çatısı.

Veri ambarları (Aytulun, 2022);

- Konu odaklıdır. Müşteri, ürün vb. varlıklar veya satış, sipariş alma, teslimat gibi olaylara odaklanmıştır. Farklı yerlerde olan veriler, belirli bir konuya uygun olarak birleştirilmiştir.
- Bütünleşiktir. Birden fazla veritabanı birbirine bağlanabilir veya İnternet gibi dış kaynaktan da veri çekilerek veritabanı ile bütünleştirilir. Dönüştürme ve normalizasyonlar yapılmıştır.
- Belirli bir zaman dilimine aittir. Örneğin 5 veya 10 yıllık bilgiler üzerinden yapılır. Her verinin bir zaman bağlamı vardır.
- Veri girişi ve çıkışına kapalıdır. Belirli bir dönemi bağladığından işlemsel veritabanından o döneme ait yeni veri çekilmez.
- Taktik ve stratejik seviyede karar vermeye yardımcı olacak veri analizi ve sorgular yapılır (Aslan burcundaki müşterilerin salı günleri hangi ürünleri aldığının belirlenmesi gibi, insan aklının yaratıcılığının sorgu ile vücut bulmuş hali gibidir).

Veri ambarları genellikle ilgili ana nesnenin etrafında inşa edilmekte, veri, tarihsel perspektiften bir bilgi üretecek şekilde saklanmaktadır. Bazen bir ürün için bazen de bir satış bölgesi için özetlenmiş bilgiler sunulabilmektedir. Veri ambarında depolanan

veri ilişkisel veri depolama şeklinde olabildiği gibi veri küpü şeklinde de olabilir. Şekil 2.4’de bir veri küpü örneği verilmiştir.



Şekil 2.4: Bir veri küpü örneği.

2.2 Veri Madenciliği İşlevleri

Veri madenciliğinin temel olarak altı farklı işlevi tanımlanmıştır (Tan vd., 2014). Bunlar; sınıflandırma (classification), kümeleme (clustering), ilişki kuralları keşfi, sıralı örüntü keşfi, regresyon ve sapma tespiti olarak bilinmektedir.

Sınıflandırma konusunda sonraki alt bölümde detaylı açıklamalar verilmiştir. Kümeleme işlevi sınıflandırmadan farklı olarak belirli bir sınıf etiketine sahip olmayan nesnelerin uygun sayıda gruplara ayrılması olarak da bilinir. Bu gruplandırma yapılırken aynı gruptaki nesnelerin benzerliği en büyük farklı gruptaki nesnelerin benzerliği en küçük olması sağlanır. Benzerlik ölçümü için cosine, jaccard, dice vb. farklı kriterler kullanılabilir. Kümeleme konusu bu çalışmanın kapsamı dışında olduğundan bu kriterlere değinilmemiştir. Kümeleme özellikle pazar ve müşteri segmentasyonu, doküman kümeleme gibi çıktıların elde edilmesinde katkı sağlamaktadır.

Birliktelik kuralları ise sıklıkla birlikte hareket eden nesnelerin keşfinde kullanılır. Perakende sektörünün başvurduğu işlevlerden biridir. Elde edilen sık desenler, güçlü

kurallar ve ilginç desenler sayesinde çeşitli pazarlama stratejilerinin geliştirilmesinde faydalanılmaktadır.

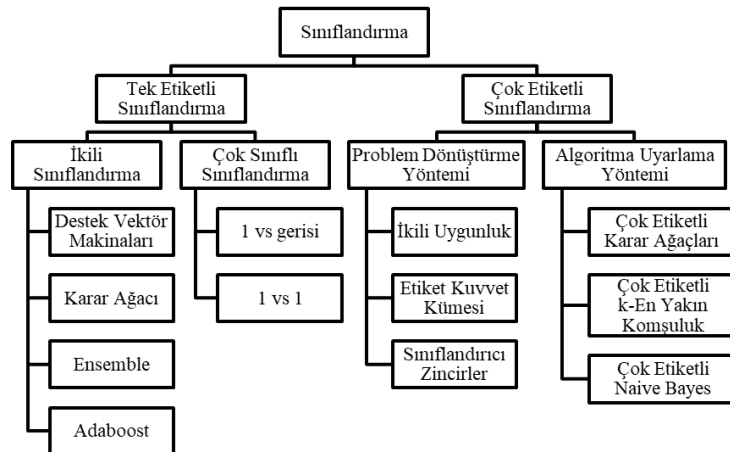
Bazen bu birlikte hareket etme durum farklı zaman aralıklarında da gerçekleşir. Bu durumda sıralı örüntü keşfine başvurulmaktadır. Müşteri alışkanlıkların tespiti gibi konularda ve/veya fiyat belirlemede kullanılmaktadır.

Veri madenciliğinin bir diğer işlevi ise zaman serileri ile ilgilidir. Temel problem bir zaman serinin gelecek değerlerinin tespitidir. Bunun için zaman serisinin oluştuğu bileşenler çeşitli istatistiksel testler yardımıyla ayrıştırılır ve her bir bileşenin gelecek değeri tahmin edilmeye çalışılır. Bu konuya yaygın olarak regresyon, arima, holt/winter, ağırlıklı ortalama, üstel düzeltme vb. tekniklerle başvurulmaktadır. Aslında yukarıda anlatılan bu işlevler çoğu zaman birlikte ele alınmaktadır. Bir problemde hem sınıflandırma hem de kümeleme algoritmaları birlikte çalışabildiği gibi zaman serisi yöntemlerinde de sınıflandırma ve kümelemede kullanılan algoritmalara başvurulabilmektedir.

2.3 Sınıflandırma Problemi

Sınıflandırma; bir veri setinde var olan örneklerle ait özniteliklerin algoritmalar tarafından anlamlandırılıp, önceden belirlenmiş ve birbirinden farklı olan sınıflardan herhangi birine atanması işlemidir.

Sınıflandırma algoritmaları; verilen eğitim veri setinden hangi örneği hangi sınıfa atayacağını öğrenir ve sonrasında geliştirdiği model ile test veri setinden sınıflarını bilmediği örnekleri doğru sınıflara atamaya çalışır.



Şekil 2.5: Sınıflandırma çeşitleri ve yöntemleri.

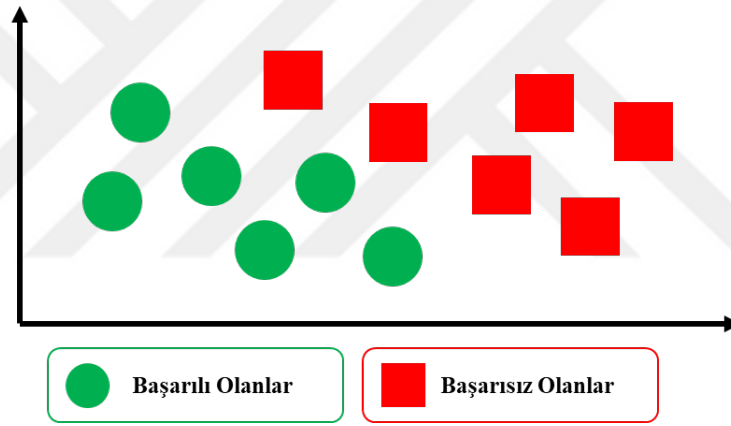
Sınıflandırma çeşitleri ve yöntemleri Şekil 2.5’de gösterilmiştir. Sınıflandırma temel olarak tek etiketli sınıflandırma ve çok etiketli sınıflandırma olmak üzere ikiye ayrılır.

2.4 Tek Etiketli Sınıflandırma

Tek etiketli sınıflandırma; her bir örneğin önceden belirlenmiş tek bir sınıf etiketindeki özniteliklerden sadece birine ait olduğu sınıflandırma çeşididir. Sınıf etiketi özniteliği sayısına bağlı olarak kendi içinde ikili sınıflandırma ve çok sınıflı sınıflandırma olmak üzere ikiye ayrılır.

2.4.1 İkili Sınıflandırma

İkili sınıflandırma; her bir örneğin önceden belirlenmiş iki sınıf etiketi özniteliğinden sadece birine ait olduğu sınıflandırma çeşididir.

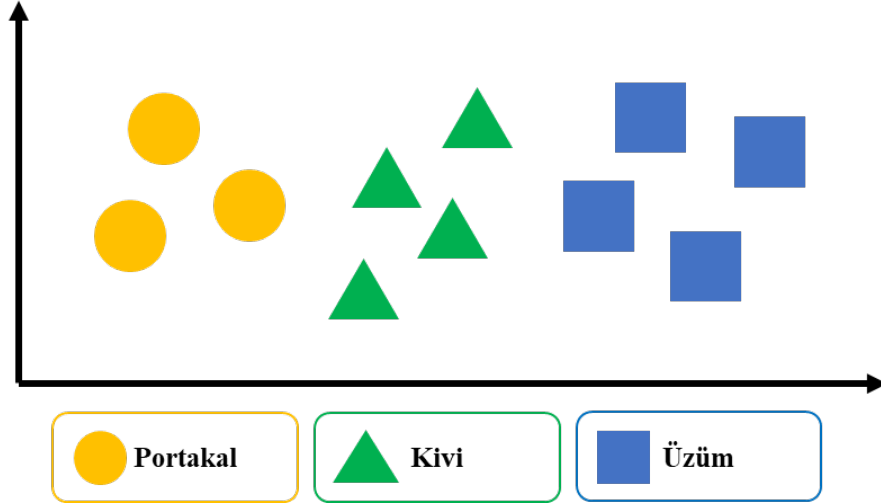


Şekil 2.6: İkili sınıflandırma örneği.

Şekil 2.6’da gösterildiği üzere bir sınava giren öğrenciler, o sınavdan başarılı (1) veya başarısız (0) olurlar. Bu tarz sınıflandırmalar ikili sınıflandırma çeşitlerindendir.

2.4.2 Çok Sınıflı Sınıflandırma

Çok sınıflı sınıflandırma; her bir örneğin önceden belirlenmiş iki veya daha fazla sınıf etiketi özniteliğinden sadece birine ait olduğu sınıflandırma çeşididir. İkili sınıflandırma ile arasındaki temel fark sınıf etiketi özniteliği sayısıdır.



Şekil 2.7: Çok sınıflı sınıflandırma örneği.

Şekil 2.7’de gösterildiği üzere bir meyve, belirlenen meyve çeşitleri arasından sadece birine ait olabilir. Aynı anda portakal ve kivi olma durumu mümkün değildir. Bu tarz sınıflandırmalar çok sınıflı sınıflandırma çeşitlerindendir.

2.5 Çok Etiketli Sınıflandırma

Çok etiketli sınıflandırma; her bir örneğin önceden belirlenmiş sınıf etiketlerinden aynı anda bir veya birden fazlasına ait olabileceği sınıflandırma çeşididir.

Örnek	Öznitelikler	Etiketler		
		Y ₁	Y ₂	Y ₃
1	X ₁	1	0	0
2	X ₂	0	1	1
3	X ₃	1	0	0
4	X ₄	1	1	1
5	X ₅	0	1	0
6	X ₆	1	1	0

Şekil 2.8: Çok etiketli veri seti örneği.

Temsili bir çok etiketli veri seti Şekil 2.8’de gösterilmiştir. Örnek eğer bir etikete ait ise “1”, değilse “0” ile gösterilir. Günümüzde örneklerin aynı anda birkaç sınıfa ait olduğu durumlar artış gösterdiği için modern uygulamalarda çok etiketli sınıflandırma yöntemlerine daha fazla ihtiyaç duyulmaktadır.

Metin sınıflandırılması, sahne sınıflandırılması gibi çok etiketli sınıflandırma problemleri en temel gerçek hayat uygulamalarındandır. Metin sınıflandırmasında her belge sanat ve beşerî bilimler gibi birden fazla konuya ait olabilmektedir (Gao vd.,

2003). Sahne sınıflandırmasında herhangi bir görüntü aynı anda deniz ve gün batımı gibi birkaç görüntü türüne ait olabilmektedir (Younes vd., 2011).

Tek etiketli veri setleri ve çok etiketli veri setleri arasındaki yapısal farklılıklardan dolayı, tek etiketli veri setleri için kullanılan yöntemler çok etiketli veri setleri için direkt olarak kullanılamamaktadır. Bu sebepten ötürü tek etiketli sınıflandırma yöntemleri ile çok etiketli veri setlerinin sınıflandırılabilmesine imkan veren bazı yöntemler geliştirilmiştir. Bu yöntemler algoritma uyarlama yöntemleri ve problem dönüştürme yöntemleri olmak üzere ikiye ayrılmaktadır.

2.5.1 Algoritma Uyarlama Yöntemi

Algoritma uyarlama yöntemleri; tek etiketli sınıflandırma yöntemlerini çok etiketli veri setleri için çalışacak şekilde uyarlayan yöntemlerdir (Grigorios ve Ioannis, 2007). Temel olarak çok etiketli karar ağaçları, çok etiketli k en yakın komşuluk ve çok etiketli naive bayes algoritması yöntemleri vardır.

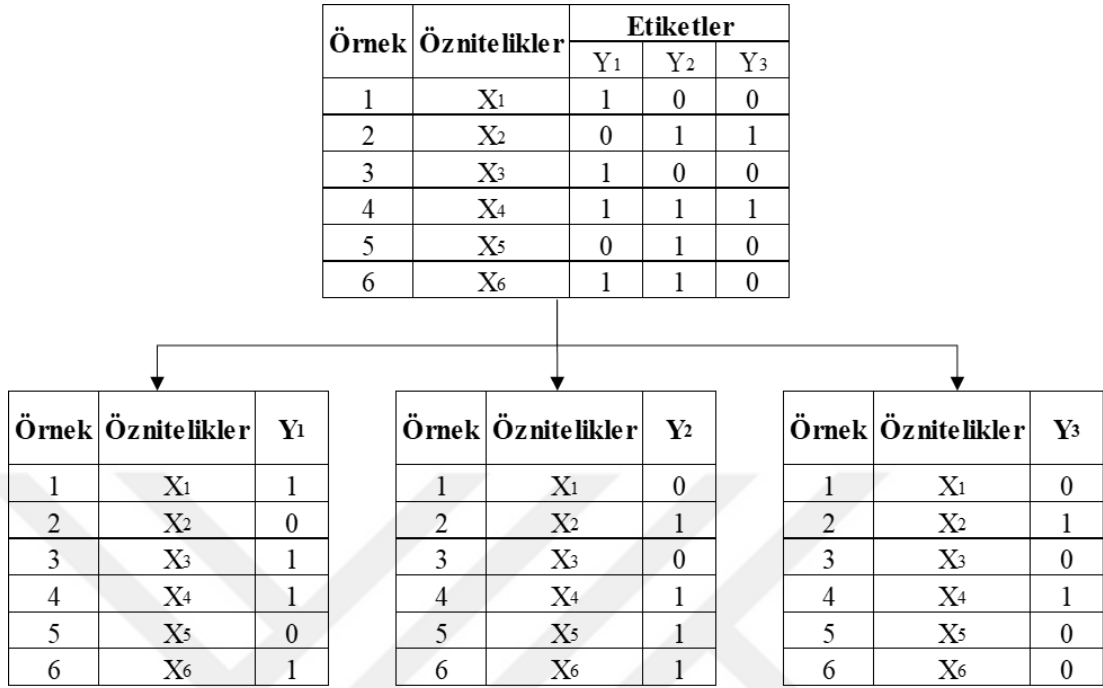
2.5.2 Problem Dönüştürme Yöntemi

Problem dönüştürme yöntemleri; çok etiketli sınıflandırma problemini bir veya birden fazla tek etiketli sınıflandırma problemine dönüştüren yöntemlerdir (Grigorios ve Ioannis, 2007). Temel olarak ikili uygunluk, etiket kuvvet kümesi ve sınıflandırıcı zincirler yöntemleri vardır.

2.5.2.1 İkili Uygunluk (Binary Relevance – BR)

Çok etiketli sınıflandırma problemi, her birinin orijinal problemdeki bir etikete karşılık geldiği bir dizi bağımsız ikili sınıflandırma problemine ayrıştırılmaktadır. Örnekler için her bir ikili sınıflandırıcıda tahminler bağımsız olarak ayrı ayrı elde edilmektedir. Elde edilen tahminler bir araya getirilip nihai çok etiketli tahmin oluşturulmaktadır (Grigorios ve Ioannis, 2007). Şekil 2.9’da her Y sınıf etiketi için problemin alt problemlere nasıl ayrıştırıldığı ve sınıflandırmanın nasıl yapıldığı tasvir edilmiştir. Bu

yöntem için her bir etiket bağımsız olarak ele alındığından etiketler arasında var olabilecek ilişkiler göz ardı edilmektedir.



Şekil 2.9: İkili uygunluk yöntemi örneği.

2.5.2.2 Etiket Kuvvet Kümesi (Label Powerset – LP)

Çok etiketli sınıflandırma problemi, orijinal problemde bulunan etiketlerin oluşturduğu tüm benzersiz kombinasyonların yeni sınıf özniteliği olarak ele alınmasıyla yeni sınıflandırma problemine, bir başka deyişle çok etiketli sınıflandırma problemi ikili veya çok sınıflı sınıflandırma problemine dönüştürülmektedir. Daha sonrasında sınıflandırmak için herhangi bir türde tek etiketli sınıflandırma yöntemi kullanılabilir (Boutell vd., 2004).

Şekil 2.10’da Y sınıf etiketlerinin oluşturduğu benzersiz kombinasyonların yeni sınıf özniteliği (Z değerleri) olarak dönüştürülmesi tasvir edilmiştir.

Örnek	Öznitelikler	Etiketler				Örnek	Öznitelikler	Etiketler
		Y ₁	Y ₂	Y ₃				
1	X ₁	1	0	0	→	1	X ₁	Z ₁
2	X ₂	0	1	1		2	X ₂	Z ₂
3	X ₃	1	0	0		3	X ₃	Z ₁
4	X ₄	1	1	1		4	X ₄	Z ₃
5	X ₅	0	1	0		5	X ₅	Z ₄
6	X ₆	1	1	0		6	X ₆	Z ₅

Şekil 2.10: Etiket kuvvet kümesi yöntemi örneği.

2.5.2.3 Sınıflandırıcı Zincirler (Classifier Chains – CC)

Çok etiketli sınıflandırma problemi, BR’da olduğu gibi her birinin orijinal problemdeki bir etikete karşılık geldiği bir dizi bağımsız ikili sınıflandırma problemine ayrıştırılır. Örnekler için her bir ikili sınıflandırıcıda tahminler bağımsız olarak ayrı ayrı elde edilmektedir. Daha sonra sınıf etiketleri için elde edilen tahmin değerleri öznitelik olarak veri setine eklenmekte ve oluşturulan yeni veri seti diğer sınıf etiketlerinin tahmininde kullanılmaktadır. Bu işlem son etiketin tahminine kadar aynı şekilde devam etmektedir (Read vd., 2009). Şekil 2.11’de sınıflandırıcı zincirler yönteminde kullanılan bir veri setinin dönüşüm işlemleri ve sınıflandırma yöntemlerinin uygulanma biçimi tasvir edilmiştir.

Örnek	Öznitelikler	Etiketler		
		Y ₁	Y ₂	Y ₃
1	X ₁	1	0	0
2	X ₂	0	1	1
3	X ₃	1	0	0
4	X ₄	1	1	1
5	X ₅	0	1	0
6	X ₆	1	1	0

Örnek	Öznitelikler	Y ₁	→	Örnek	Öznitelikler		Y ₂	→	Örnek	Öznitelikler			Y ₃
					X	Y ₁ '				X	Y ₁ '	Y ₂ '	
1	X ₁	1		1	X ₁	1	0		1	X ₁	1	0	0
2	X ₂	0		2	X ₂	0	1		2	X ₂	0	1	1
3	X ₃	1		3	X ₃	1	0		3	X ₃	1	0	0
4	X ₄	1		4	X ₄	1	1		4	X ₄	1	1	1
5	X ₅	0		5	X ₅	0	1		5	X ₅	0	1	0
6	X ₆	1		6	X ₆	1	1		6	X ₆	1	1	0

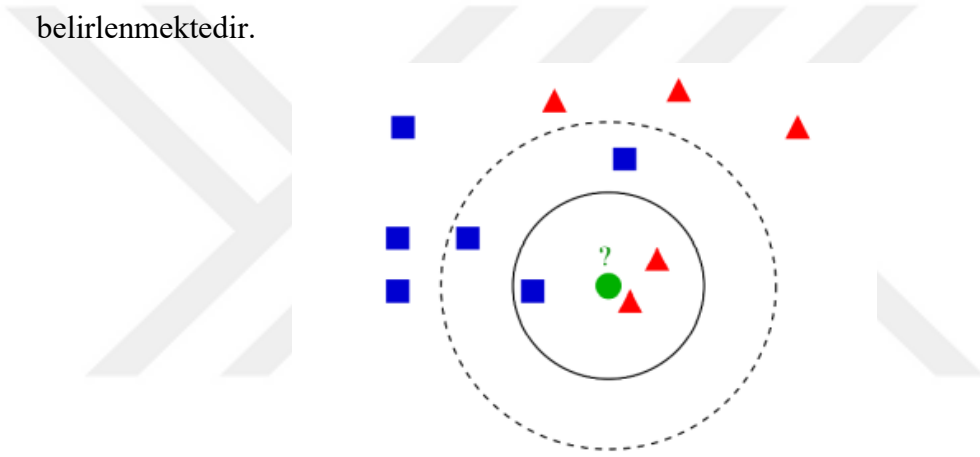
Şekil 2.11: Sınıflandırıcı zincirler yöntemi örneği.

2.6 Sınıflandırma Algoritmaları

Bu bölümde çalışmada kullanılan sınıflandırma algoritmalarından bahsedilmiştir.

2.6.1 k En Yakın Komşuluk Algoritması (KNN)

k en yakın komşuluk algoritması (KNN) özellikle küçük veri setine sahip problemlerde iyi sonuçlar veren basit bir kaba kuvvet algoritmasıdır. Şekil 2.8’de KNN algoritmasının örneği gösterilmiştir. KNN benzer örneklerin birbirlerine yakın olduğu varsayımı ile sınıflandırma yapmaktadır. Sınıflandırma için değerlendirilecek olan komşu sayısı k olup, modelin örneği iyi sınıflandırabilmesi için en uygun değerin seçilmesi gerekmektedir. k değeri, eşitlik olmaması için tek tam sayı olarak belirlenmektedir.



Şekil 2.12: k en yakın komşu algoritması örneği.

Bu yöntem de sınıflandırılmamış bir örnek sınıf etiketi belli olan örneklerden hangisine en yakın ise o sınıf grubuna atmaktadır. KNN algoritması uygulama basamakları:

- 1- Başlangıç k değeri belirlenir.
- 2- Veri kümesinde bulunan bütün örneklerle arasındaki uzaklık çeşitli fonksiyonlar yardımıyla hesaplanır.
- 3- En yakın k adet komşu tespit edilir.
- 4- k adet komşudan fazla olan sınıfa atanır.

KNN algoritması sınıflandırma yöntemi olarak çok uzun zamandır kullanılmaktadır. Basittir ve sınıflandırma hatası düşüktür. Dezavantaj olarak eğitim verileri için donanımsal olarak büyük hafızaya ihtiyaç duymaktadır (Cover ve Hart, 1967).

Uzaklık ölçüm fonksiyonları olarak Öklid (2.1), Manhattan (2.2), Minkowski (2.3) ve Chebyshev (2.4) kullanılabilir.

$$d_{öklid}(x_1, x_2) = \sqrt{\sum_{i=1}^n (x_{1i} - x_{2i})^2} \quad (2.1)$$

$$d_{minkowski}(x_1, x_2) = \sqrt[n]{\sum_{i=1}^n |x_{1i} - x_{2i}|^n} \quad (2.2)$$

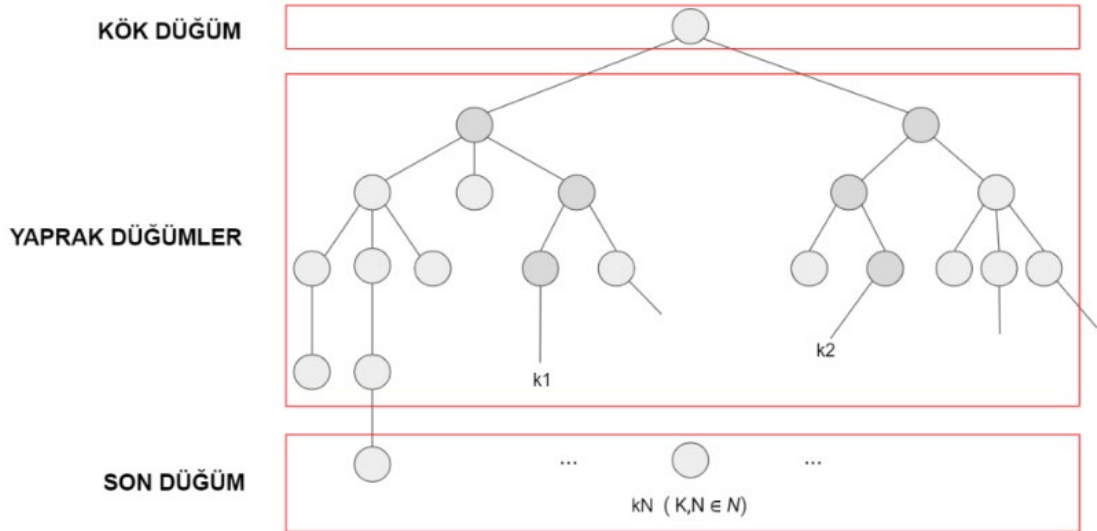
$$d_{manhattan}(x_1, x_2) = \sum_{i=1}^n |x_{1i} - x_{2i}| \quad (2.3)$$

$$d_{chebyshev}(x_1, x_2) = \max_i |x_{1i} - x_{2i}| \quad (2.4)$$

KNN algoritması içerisinde farklı ölçütlerde kullanılarak doğru sınıflandırma tahminleri yapabildiğini gösteren çalışmalar da mevcuttur (Weinberger ve Saul, 2006).

2.6.2 Karar Ağacı Algoritması (DET)

Karar ağacı algoritması (DET), Şekil 2.13'deki gibi kök düğümden yaprak düğümlere doğru ilerleyen yönlü bir ağaç yapısı oluşturarak sınıflandırma yapabilen ve hem regresyon hem de sınıflandırma problemleri için kullanılabilen bir algoritmadır.



Şekil 2.13: Karar ağacı yapısı örneği.

Hem kategorik hem sayısal verilere uygulanabilen yöntem uygulaması kolay ve yorumlanması basittir. Gürültülü, eksik ve aşırı değerlere sahip veri setlerinde sıkça

tercih edilen bir yöntemdir. C4.5, C5.0 gibi birden fazla geliştirilmiş karar ağacı algoritması literatürde mevcuttur.

2.6.3 Naive Bayes Algoritması (NAB)

Naive bayes algoritması (NAB), temeli bayes teoremine dayanan bir sınıflandırma algoritmasıdır. Bu yöntemde örneklerin hangi sınıfa ait olabileceğine dair olasılıklar hesaplanmakta ve örnek en yüksek olasılığa sahip olan sınıfa atanmaktadır.

Dengesiz veri kümelerinin sınıflandırılmasında da geçerliliği olan bir yöntemdir. Bu yöntemde, her bir değişkenin hangi sınıfa ait olacağı birbirinden bağımsızdır. Bir değişkenin olasılığı diğer bir değişkenin olasılığını etkilememektedir (Mukherjee ve Sharma, 2012).

Hesaplama kolaylığı ve basitliği sebebiyle sınıflandırma problemlerinin çözümünde sıkça kullanılmaktadır. Denklem 2.5’de bayes hesaplaması verilmiştir.

$$P(b|a) = P(b)P(a|b)/P(a) \quad (2.5)$$

Denklem 2.5’de a olayının gerçekleşme ihtimali $P(a)$ iken. b olayının gerçekleşme ihtimali ise $P(b)$ ’dir. $P(b|a)$ ile ifade edilmek istenen, a olayı gerçekleştiğinde b olayının olma olasılığıdır. Naive Bayes Sınıflandırıcısı, özniteliklerin sınıfa göre koşullu olarak bağımsız olduğu varsaymaktadır.

Denklem 2.6 ve 2.7’de a_i , a ’daki i . özniteliği, m öznitelik sayısını, n sınıf sayısını ve c_i i . sınıfı temsil etmektedir (Claude ve Webb, 2011).

$$P(a|b) = \prod_{i=1}^m P(a_i|b) \quad (2.6)$$

$$P(a) = \prod_{i=1}^n P(c_i)P(a|c_i) \quad (2.7)$$

2.6.4 Lojistik Regresyon Algoritması (LOG)

Lojistik regresyon algoritması (LOG), bir veya birden fazla bağımsız değişkenin ikili bir sonuç değişkeni üzerindeki etkisini değerlendirmek, hedef değişken y ile bir dizi değişken olan x arasındaki ilişkiyi kullanarak sınıflandırma yapan bir algoritmadır.

2.6.5 Genelleştirilmiş Doğrusal Model Algoritması (GLM)

Nelder ve Wedderburn tarafından 1972 yılında tanıtılan genelleştirilmiş doğrusal model algoritması (GLM), üssel dağılım ailesinin üyelerinden birinin, araştırmacı tarafından bir yanıt dağılımı olarak seçimine müsaade eden doğrusal ve doğrusal olmayan regresyon modellerin bir birleşimidir.

Normal, Binom, Poisson, Üstel, Gamma, Ters Normal ve Negatif Binom dağılımları üssel ailenin birer üyesidir. GLM ile normallik ve sabit varyans varsayımı gerekmemektedir ancak tüm üssel aile üyelerinde varyans ortalamasının bir fonksiyonudur.

2.6.6 Derin Öğrenme Algoritması (DEL)

Derin öğrenme algoritması (DEL), birbiri ardına sıralanmış katmanlarda benzersiz girdi verilerinden başlanarak veriler sırasıyla işlenirken gittikçe artan şekilde daha kullanılabilir gösterimler elde edebilen makine öğrenmesi yöntemidir.

Katmanlı yapısı, sinir ağı olarak tanımlanmaktadır. Veri, sinir ağları ve birbiri ardına sıralanmış katmanları olan model ile öğrenilmektedir. Katman sayısı arttıkça derinlik artmakta ve veriden daha fazla öğrenme gerçekleşmektedir.

2.6.7 Rassal Orman Algoritması (RAF)

Karar ağacı algoritmaları üzerinden geliştirilen rassal orman algoritması (RAF) bir topluluk öğrenme yöntemidir. Çok sayıda, rassal ve birbirinden bağımsız olan karar ağaçlarını, güçlü bir sınıflandırma işlemi yapmak için kullanan bir algoritmadır. Şekil 2.14'deki gibi kullanılan eğitim veri seti sayesinde oluşturulan sınıflandırıcı karar

ağaçlarının bir araya gelmesiyle karar ormanı oluşur. Oluşturulan karar ormanında oy çokluğuna göre sınıflandırma gerçekleşir (Breiman, 2001).

Şekil 2.14: Rassal orman yapısı örneği.

2.6.8 Gradyan Güçlendirilmiş Ağaçlar Algoritması (GBT)

3. BAŞARI DEĞERLENDİRME ÖLÇÜTLERİ

Sınıflandırma yöntemlerini ve algoritmalarını değerlendirmek ve hangisinin seçileceğine karar vermek için değerlendirme ölçütlerine ihtiyaç duyulmaktadır (Oprea ve Ti, 2014). Bu kısımda sınıflandırma algoritmaları için kullanılan değerlendirme ölçütlerinden bahsedilmektedir.

3.1 Karmaşıklık Matrisi (Confusion Matrix)

Tablo 3.1: Karmaşıklık matrisi.

Karmaşıklık Matrisi		Gerçek Sonuçlar	
		Pozitif (1)	Negatif (0)
Tahmin Edilen Sonuçlar	Pozitif (1)	TP	FP
	Negatif (0)	FN	TN

Tablo 3.1’de bulunan karmaşıklık matrisi, sınıflandırma yapılırken gerçekleşen tahminlerin ne kadar başarılı olup olmadığını göstermektedir. Karmaşıklık matrisi içinde bulunan tahmin değerlerini örnek vererek kısaca açıklanacak olursa;

- True Positive (TP): Eğer bir kişinin aşı olduğunu (1) tahmin ettiysek ve o kişi gerçekte aşı olduysa (1) ise bu değeri almaktadır.
- True Negative (TN): Eğer bir kişinin aşı olmadığını (0) tahmin ettiysek ve o kişi gerçekte aşı olmadıysa (0) bu değeri almaktadır.
- False Positive (FP): Eğer bir kişinin aşı olduğunu (1) tahmin ettiysek ve o kişi gerçekte aşı olmadıysa (0) bu değeri almaktadır.
- False Negative (FN): Eğer bir kişinin aşı olmadığını (0) tahmin ettiysek ve o kişi gerçekte aşı olduysa (1) ise bu değeri almaktadır.

3.2 Doğruluk (Accuracy)

Doğruluk değeri; doğru sınıflandırılanların yüzdesidir. Denklem 3.1 ile hesaplanmaktadır (Eyupoglu vd., 2018).

$$\text{Doğruluk (Accuracy)} = \frac{TP+TN}{TP+FP+TN+FN} \quad (3.1)$$

Doğru tahminlerin, toplam tahmine oranı şeklinde de tanımlanmaktadır. Sınıflandırmanın yüzdesel olarak doğruluk başarısını göstermektedir. Veri kümelerinin dengesiz dağıldığı durumlarda bu ölçüt tek başına yeterli olmamaktadır.

3.3 Kesinlik (Precision)

Kesinlik değeri; pozitif olarak tahmin edilen değerlerin gerçekte yüzdesel olarak ne kadarının pozitif olduğunu göstermektedir. Denklem 3.2 ile hesaplanmaktadır.

$$\text{Kesinlik (Precision)} = \frac{TP}{TP+FP} \quad (3.2)$$

Tüm sınıflardan, doğru olarak ne kadar tahmin edildiğinin bir ölçüsüdür. Mümkün olduğu kadar yüksek olmalıdır. Pozitif Tahmin Edici Değer olarak da bilinmektedir.

3.4 Duyarlılık (Recall)

Duyarlılık değeri; pozitif olarak tahmin edilmesi gereken örneklerin ne kadarını pozitif olarak tahmin edildiğini gösteren bir değerdir. Denklem 3.3 ile hesaplanmaktadır (Vuk ve Curk, 2006).

$$\text{Duyarlılık (Recall)} = \frac{TP}{TP+FN} \quad (3.3)$$

Ölçütlerin önemi, modelin hangi amaçla geliştirildiğine göre değişmektedir. Örneğin aşı olacak kişileri tespit etmeye çalıştığımız bir modelde, duyarlılık değeri çok önemlidir ve oldukça yüksek olmalıdır.

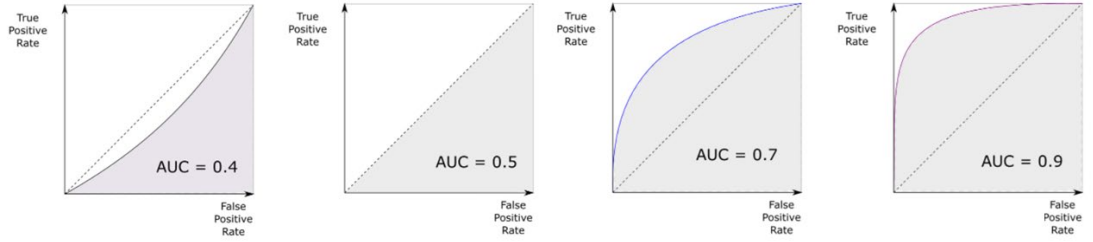
3.5 F Puanı (F Score)

F Puanı, kesinlik ve duyarlılık değerlerinin harmonik ortalamasını göstermektedir. F Puanına ihtiyaç duyulmasındaki en önemli sebep ise dengesiz dağılan veri kümelerinde hatalı bir model seçimi yapılmaması için mevcut ölçütlerin yetersiz kalmasıdır. Buna ek olarak bütün hata maliyetlerini de kapsayacak bir ölçme metriğine ihtiyaç duyulduğu için F Puanı değerlendirme için oldukça uygun bir ölçütüdür (Chimieski ve Fagundes, 2013). Denklem 3.4 ile hesaplanmaktadır.

$$F \text{ Puanı (F Score)} = 2 * \frac{\text{Kesinlik} * \text{Duyarlılık}}{\text{Kesinlik} + \text{Duyarlılık}} \quad (3.4)$$

3.6 Eğri Altında Kalan Alan (Area Under the Curve - AUC)

Sınıflandırma algoritmalarının tüm olası değerler üzerinde performansını özet olarak ortaya koymak için tercih edilen bir grafikdir. Alıcı işletim karakteristiği (Receiver Operating Characteristic – ROC) eğrisinin altında kalan alan (Area Under the Curve - AUC), iki sınıf arasında ne ölçüde kuvvetli bir ayırım yapılabileceğini gösterir. Şekil 3.1’deki gibi alan arttıkça ayırım yeteneği artar (Akca, 2020)s.



Şekil 3.1 ROC AUC grafikleri.

4. UYGULAMA

Çalışmada ele alınan problemi çözmek için önerilen yöntem ve algoritmalar RapidMiner Studio (Versiyon 9.10.001) ortamında; Windows 10 işletim sisteminde çalışan, M2 SSD sabit disk, 16 GB RAM ve 2.60 GHz 12 çekirdekli Intel işlemciye sahip bir bilgisayar ile gerçekleştirilmiştir.

4.1 Veri Seti Özellikleri

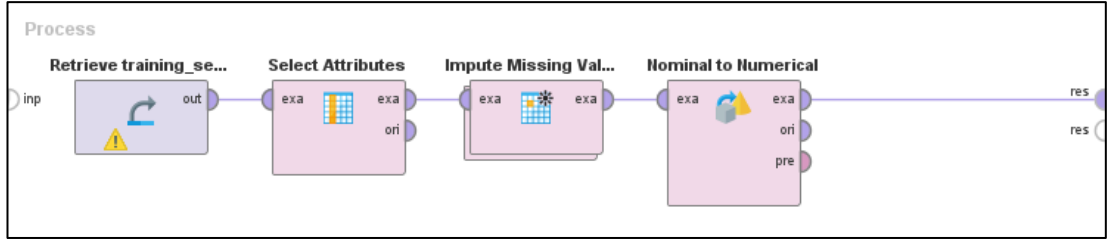
Bu çalışmada kullanılan veri seti Amerika Birleşik Devletleri tarafından 2009 yılı sonu ve 2010 yılı başlarında yapılan Ulusal H1N1 Grip aşısı anketinden drivendata.org sitesi aracılığıyla elde edilmiştir. Bu ankette katılımcılara kendileri hakkındaki sorularla birlikte H1N1 grip aşısı ve mevsimsel grip aşısı olup olmadıkları sorulmuştur. Bu sorulara ek olarak sosyal, ekonomik ve demografik bilgileri, hastalık riskleri ve aşıların etkinlikleri hakkındaki görüşleri ve hastalık bulaştırmayı azaltmaya yönelik davranışları sorulmuştur. Veri seti 26.707 adet kayıttan ve 38 adet öznitelikten oluşmaktadır. Veri setinde bulunan özniteliklere ait bilgiler Ek-1’de gösterilmiştir.

4.2 Veri Ön İşleme

Veri seti incelendiğinde, Tablo 4.1’de adı geçen öznitelik değerlerinin veri setinin büyük çoğunluğunda eksik olduğu tespit edilmiştir. Bu özniteliklerin etkinliğinin düşük olması sebebiyle Şekil 4.1’deki “Select Attributes” operatörü ile veri setinden çıkartılmıştır.

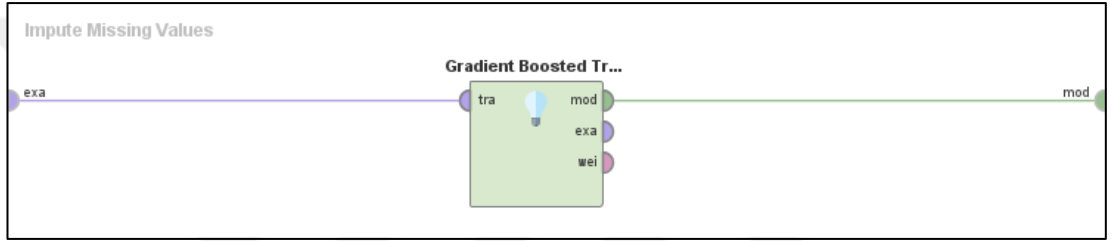
Tablo 4.1: Çok eksik değer içeren öznitelikler.

Öznitelik Adı	Eksik Veri Miktarı
employment occupation	13.470
empoyment industry	13.330
health insurancet	12.270



Şekil 4.1: RapidMiner Studio veri ön işleme süreç akışı.

Mevcut veri setinde geriye kalan eksik değerler Şekil 4.1’deki “Impute Missing Values” operatörü içinde, Şekil 4.2’deki “Gradient Boosted Trees” operatörü kullanılarak tamamlanmıştır.



Şekil 4.2: Impute Missing Values operatörü süreci.

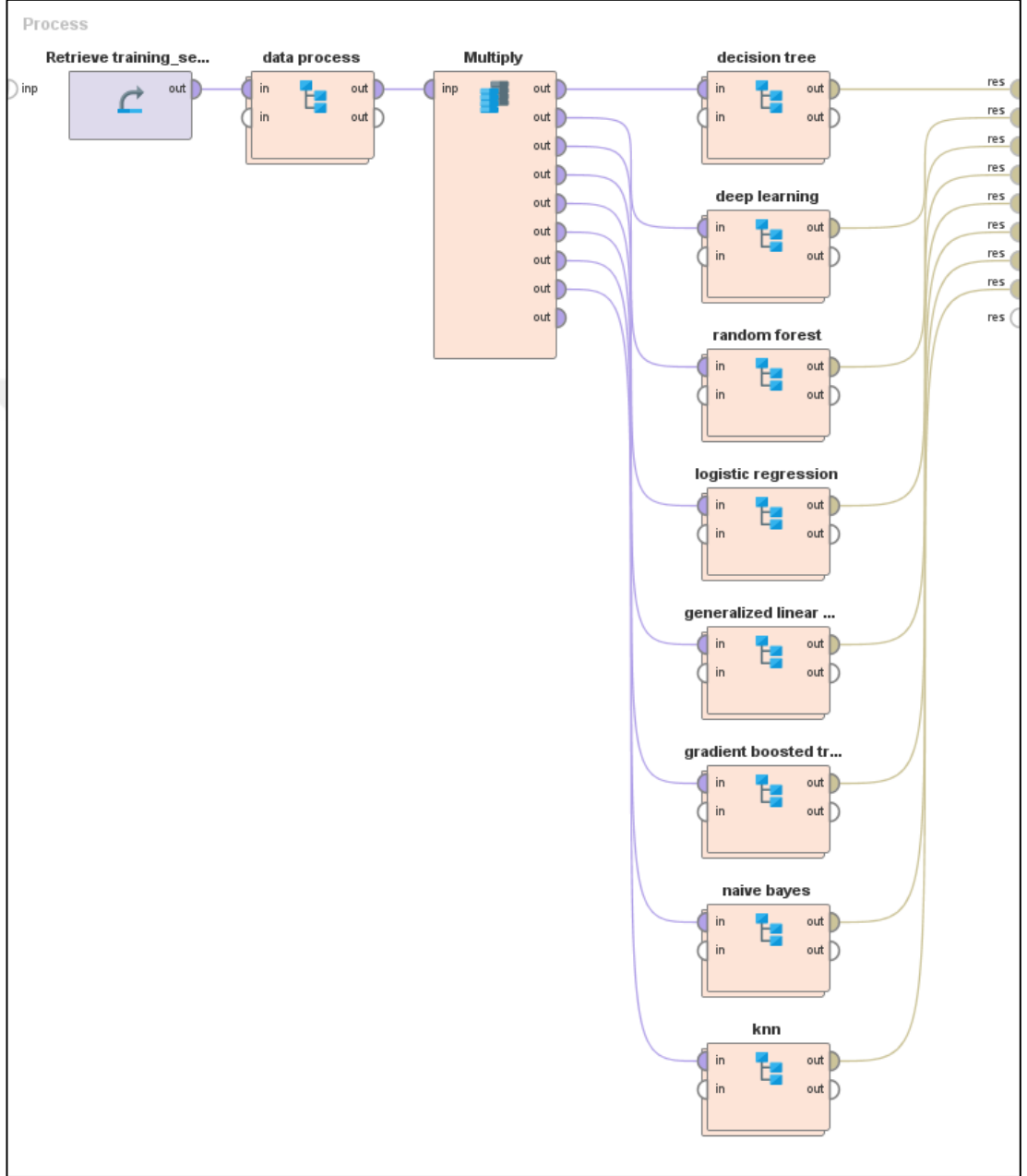
Mevcut veri setinde Tablo 4.2’de yer alan öznitelikler nominal değerlere sahip oldukları için bu değerlerin sayısal değerlere dönüştürülmesi gerekmektedir. Kukla değişkenler kodlaması ile Şekil 4.2’deki “Nominal to Numerical” operatörü kullanılarak değerlerin dönüştürülmesi sağlanmıştır.

Tablo 4.2: Nominal değer içeren öznitelikler.

Öznitelik Adı	Öznitelik Değerleri
age_group	18-34, 35-44, 45-54, 55-64, 65+
census_msa	Non-MSA, MSA-Not Principle City, MSA-Principle City
education	college graduate, some college, 12 years, <12 years
employment_status	employed, unemployed
hhs_geo_region	gizlenmiş 10 adet bölge
income_poverty	below poverty, <= \$75.000 above poverty, > \$75.000
marital_status	married, not married
race	white, black, hispanic, other or multiple
rent_or_own	rent, own
sex	female, male

4.3 İkili Uygunluk Yöntemi Uygulaması

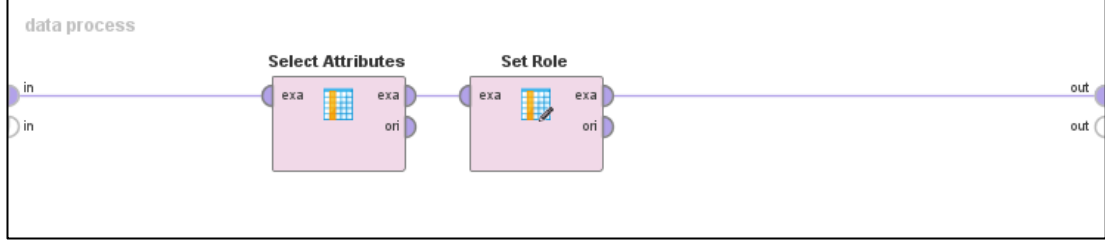
İkili Uygunluk Yöntemi için veri seti her iki etiketten sadece birini içerecek şekilde ikiye ayrılmış ve ikili sınıflandırma yapılmıştır.



Şekil 4.3: RapidMiner Studio ikili uygunluk yöntemi süreç akışı.

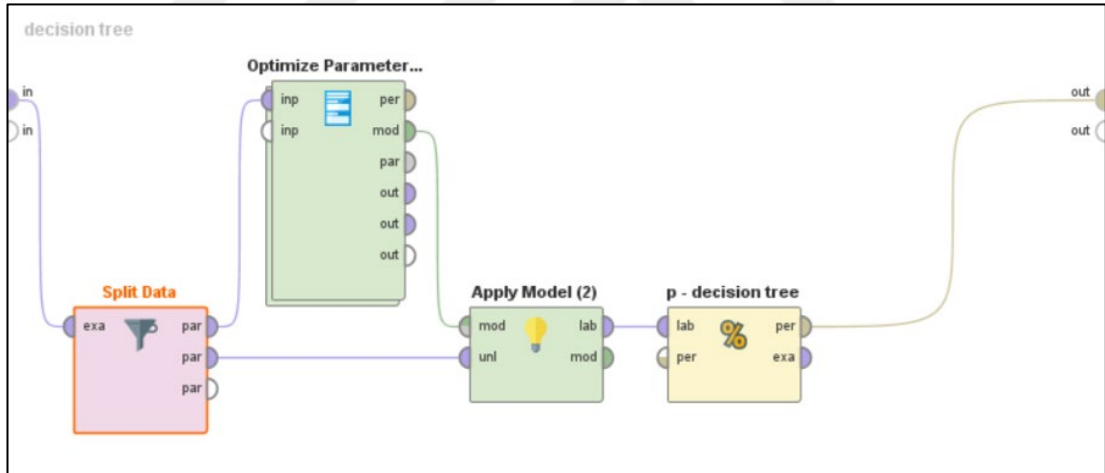
RapidMiner Studio’da oluşturulan uygulamanın arayüzü Şekil 4.3’de gösterilmiştir. “Retrieve” operatörü kullanılarak veri ön işleme safhasında oluşturulan veri setine ulaşılmaktadır. Daha sonrasında “data process” operatörünün altında Şekil 4.4’de gösterildiği üzere; “Select Attributes” operatörü ile tahmini yapılmayacak sınıf etiketi veri setinden çıkartılmış; “Set Role” operatörü ile tahmini yapılacak sınıf etiketi

“etiket” olarak işaretlenmiştir. Şekil 4.3’deki “Multiply” operatörü ile tahmin yapılmak için kullanılacak algoritmaların olduğu operatörlere aynı veri seti kopyalanarak gönderilmiştir.

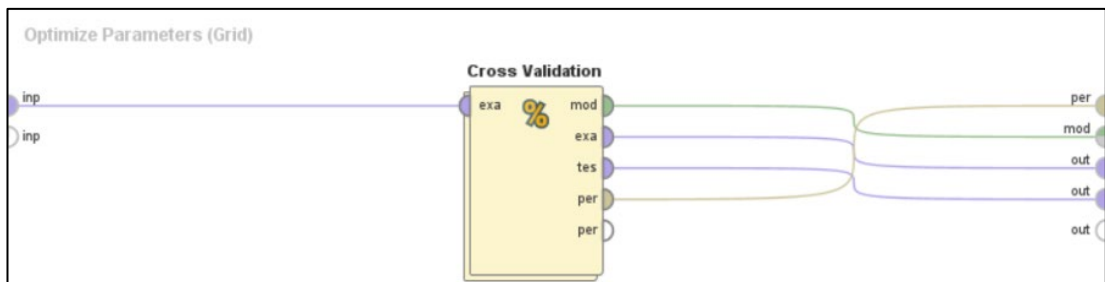


Şekil 4.4: Data process operatörü süreci.

Karar ağacı algoritmasının RapipMiner Studio’daki süreci Şekil 4.5’de gösterilmiştir. “Split Data” operatörü ile hold-out metodu kullanılarak veri setinin %70’i eğitim, %30’u test verisi olarak ikiye ayrılmaktadır. “Optimize Parameters” operatörü Şekil 4.6’de görülen “Cross Validation” operatörünü çapraz doğrulama metodunun k katsayısını ve örneklem tipini optimize ederek en iyi sonucuna ulaşmaktadır.

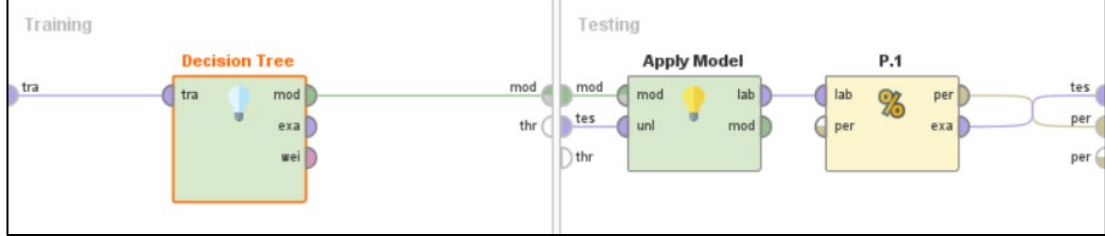


Şekil 4.5: Karar ağacı algoritması operatörü süreci.



Şekil 4.6: Karar ağacı algoritması “Optimize Parameters” operatörü süreci.

Şekil 4.7’de ise “Cross Validation” operatörünün altında DET algoritmasının modellenmesi gösterilmiştir. Diğer sınıflandırma algoritmaları da benzer şekillerde tasarlanmıştır. Sadece DET ve RAF algoritmalarında “Optimize Parameters” operatörü kullanılmış, diğerlerinde kullanılmamıştır.



Şekil 4.7: Karar ağacı algoritması “Cross Validation” operatörü süreci.

Sınıflandırma algoritmaları ve performans değerleri H1N1 grip aşısı için Tablo 4.3’de ve mevsimsel grip aşısı için Tablo 4.4’de gösterilmiştir. H1N1 grip aşısı tahmini için GBT sınıflandırma algoritması Tablo 4.3’te gösterildiği üzere 0,587 F Puanı ile en iyi sonucu verirken, mevsimsel grip aşısı tahmini için de GBT sınıflandırma algoritması Tablo 4.4’te gösterildiği üzere 0,768 F Puanı ile en iyi sonucu vermiştir.

Tablo 4.3: H1N1 grip aşısı olma tahmini performansı.

Algoritma	Doğruluk (%)	Kesinlik (%)	Duyarlılık (%)	AUC	F Puanı
DET	81,96	59,40	47,71	0,784	0,529
DEL	82,05	57,60	58,75	0,819	0,582
GLM	83,71	69,56	41,48	0,835	0,520
GBT	80,60	53,59	64,98	0,824	0,587
LOG	83,97	69,03	44,54	0,839	0,541
NAB	78,16	48,86	60,22	0,792	0,539
RAF	82,96	75,19	29,55	0,835	0,424
KNN	80,20	56,99	27,79	0,708	0,374

Tablo 4.4: Mevsimsel grip aşısı olma tahmini performansı.

Algoritma	Doğruluk (%)	Kesinlik (%)	Duyarlılık (%)	AUC	F Puanı
DET	75,92	73,82	74,83	0,812	0,743
DEL	75,96	71,02	81,72	0,838	0,760
GLM	78,52	78,19	74,69	0,857	0,764
GBT	77,62	74,22	79,57	0,851	0,768
LOG	78,15	77,46	74,83	0,857	0,761
NAB	73,19	69,69	75,07	0,799	0,723
RAF	78,52	78,58	74,05	0,861	0,762
KNN	70,96	69,49	67,05	0,766	0,682

4.4 Etiket Kuvvet Kümesi Yöntemi Uygulaması

Etiket Kuvvet Kümesi Yöntemi için mevcut sınıf etiketlerinin olası bütün kombinasyonları göz önünde bulundurularak yeni sınıf etiketleri oluşturulmuştur. Buna göre H1N1 grip aşısı ve mevsimsel grip aşısı olanlar “4”, sadece H1N1 grip aşısı olanlar “1”, sadece mevsimsel grip aşısı olanlar “3”, hiçbirini olmayanlar “2” sınıf etiketi ile işaretlenmiştir. Etiket Kuvvet Kümesi Yöntemi ile elde edilen sonuçlar, etiketlerin sayılarına göre ağırlıkları ile çarpılıp Tablo 4.5’te bulunan tahmin sonuçları elde edilmiştir.

Tablo 4.5: Etiket kuvvet kümesi yöntemi tahmin performansı.

Algoritma	Doğruluk (%)	Kesinlik (%)	Duyarlılık (%)	F Puanı
DET	64,71	62,88	64,69	0,633
DEL	67,34	65,98	67,31	0,658
GLM	68,58	67,14	68,55	0,668
GBT	67,66	65,87	67,63	0,659
NAB	56,15	59,87	56,13	0,577
RAF	67,59	64,68	67,56	0,648
KNN	59,90	56,68	59,87	0,575

Etiket Kuvvet Kümesi Yöntemi için en iyi performansı Genelleştirilmiş Doğrusal Model algoritması Tablo 4.5’te görüldüğü üzere 0,668 F Puanı değeriyle gerçekleştirmiştir.

4.5 Sınıflandırıcı Zincirler Yöntemi Uygulaması

İkili Uygunluk Yöntemi Uygulaması kısmında H1N1 Grip Aşısı sınıf etiketi için Tablo 4.3’te, Mevsimsel Grip Aşısı sınıf etiketi için Tablo 4.4’te gösterildiği üzere en iyi sınıflandırma algoritması GBT algoritması olarak tespit edilmiştir. Her iki sınıf etiketi için GBT algoritması kullanılarak elde edilen tahmin değerleri, sınıflandırıcı zincirler yönetimini uygulamak için veri setine özellik olarak eklenmiştir.

H1N1 Grip Aşısı sınıf etiketi tahmini için Mevsimsel Grip Aşısı tahmin değerlerinin özellik olarak kullanıldığı Sınıflandırıcı Zincirler Yöntemi tahmin sonuçları Tablo 4.6’da, Mevsimsel Grip Aşısı sınıf etiketinin tahmini için H1N1 Grip Aşısı tahmin

değerlerinin özellik olarak kullanıldığı Sınıflandırıcı Zincirler Yöntemi tahmin sonuçları Tablo 4.7’de gösterilmiştir.

**Tablo 4.6: sınıflandırıcı zincirler yöntemi
h1n1 grip aşısı olma tahmini performansı.**

Algoritma	Doğruluk (%)	Kesinlik (%)	Duyarlılık (%)	AUC	F Puanı
DET	82,99	64,75	43,71	0,815	0,522
DEL	81,76	56,88	58,52	0,812	0,577
GLM	83,89	69,63	42,83	0,835	0,530
GBT	80,80	53,98	65,28	0,825	0,591
LOG	84,11	69,27	45,30	0,839	0,548
NAB	77,16	47,19	63,04	0,793	0,540
RAF	83,11	77,74	28,73	0,835	0,420
KNN	75,31	23,36	7,11	0,518	0,109

**Tablo 4.7: Sınıflandırıcı zincirler yöntemi
mevsimsel grip aşısı olma tahmini performansı.**

Algoritma	Doğruluk (%)	Kesinlik (%)	Duyarlılık (%)	AUC	F Puanı
DET	76,70	76,87	71,45	0,815	0,741
DEL	76,68	72,52	80,38	0,846	0,762
GLM	78,52	78,26	74,58	0,858	0,764
GBT	77,56	73,98	79,89	0,851	0,768
LOG	78,16	77,52	74,77	0,857	0,761
NAB	73,30	69,96	74,75	0,803	0,723
RAF	78,16	78,29	73,46	0,859	0,758
KNN	50,67	46,88	44,75	0,513	0,458

4.6 Sonuç ve Bulgular

Elde edilen bütün sonuçların en iyi değerleri Tablo 4.8’de gösterilmiştir. GBT algoritması genel olarak bütün yöntemlerde, sınıf etiketi tahmininde başarılı olmuştur.

Tablo 4.8: Karşılaştırma tablosu.

Yöntem	Etiket	Algoritma	Doğruluk (%)	Kesinlik (%)	Duyarlılık (%)	AUC	F Puanı
İkili Uygunluk	Mevsimsel Grip Aşısı	GBT	77,62	74,22	79,57	0,851	0,768
	H1N1 Grip Aşısı	GBT	80,60	53,59	65,98	0,824	0,587
Etiket Kuvvet Kümesi	-	GLM	68,58	67,14	68,55	-	0,668
Sınıflandırıcı Zincirler	Mevsimsel Grip Aşısı	GBT	77,56	73,98	79,89	0,851	0,768
	H1N1 Grip Aşısı	GBT	80,80	53,98	65,28	0,825	0,591

5. SONUÇ VE ÖNERİLER

Yakın zamanda dünyayı etki altına alan COVID-19 salgınının etkileri hala devam etmektedir. Bu salgın ile beraber tüm dünyada bir kez daha anlaşılmıştır ki; salgınlarda etkili bir aşının en kısa zamanda geliştirilmesi ve toplumun yeterli kısmına aşılamanın yapılması büyük bir önem arz etmektedir.

Bu çalışmada aşılama öncesi yapılan anket çalışmaları sonucunda elde edilen bilgiler veri madenciliği yöntemleri kullanılarak işlenmiştir. Çalışmada yapılan deneyler ile aşılama bölgelerine gönderilecek olan aşı tedarik miktarlarında gerçeğe yakın tahminlerin yapılması için gerekli çözümler sunulmuştur.

Uygulanan problem dönüştürme yöntemlerinden Sınıflandırıcı Zincirleri yöntemi, H1N1 grip aşısı tahmini için İkili Uygunluk yöntemi ile aynı başarı ölçütlerine sahipken, mevsimsel grip aşısı tahmini için daha iyi sonuçlar vermiştir. Bu sebeple bir etiketin tahmin değerlerinin diğer etiketin tahmininde kullanılmasının performansı arttırabileceği en azından bu veri seti için söylenebilir.

Etiket Kuvvet Kümesi yöntemi problemi 4 sınıfa ayırmıştır. Sınıfların veri kümesi içinde dengesiz dağılması sebebiyle yeterli öğrenme gerçekleşmemesinden dolayı iyi performans göstermemiştir.

Çalışmanın literatüre ana katkısı olarak hedeflenen, kullanılamayan aşı miktarının en aza indirilmesi ve toplumun aşılama miktarının maksimize edilmesi amacımıza ulaşılmıştır. Böylelikle çalışmamızın gelecekteki halk sağlığı çalışmaları için yol gösterebileceği net olarak ortaya konulmuştur.

Bundan sonraki çalışmalarda iki etiketli veri setleri dışında ikiden fazla sınıf etiketi barındıran veri setleri incelenerek daha gerçekçi sonuçlara erişmek mümkün olabilecektir. Ayrıca literatürde tanımlı yaklaşımlar dışında bulunan performans ölçütlerinin değerlerini yükseltebilecek yeni algoritmaların da geliştirilebileceği değerlendirilmektedir.

KAYNAKÇA

- Akca, M. F. 2020. "Sınıflandırma Problemlerindeki Metrikler". Tarihinde adresinden erişildi <https://medium.com/deep-learning-turkiye/sınıflandırma-problemlerinde-ki-metrikler-33ee5f30f8eb>
- Aytulun, S. K. 2022. "Veri Madenciliği Ders Notları 2022".
- Boutell, M. R., Luo, J., Shen, X., Brown, C. M. 2004. "Learning multi-label scene classification", 37, 1757–1771.
- Breiman, L. 2001. "Random Forests". Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 12343 LNCS.
- Cerrito, P. B. 2006. *"Introduction to Data Mining using SAS Enterprise Miner"*.
- Çevik, V. A. 2021. "The Most Comprehensive Cold Chain Application in History : Challenges in Logistics of Covid-19 Vaccine". Üsküdar Üniversitesi Sosyal Bilimler Dergisi, 2021(Mayıs), 155–204.
- Chimieski, B., Fagundes, R. 2013. "Association and Classification Data Mining Algorithms Comparison over Medical Datasets". Journal of Health Informatics, 5(2), 44–51. Tarihinde adresinden erişildi <http://www.jhi-sbis.saude.ws/ojs-jhi/index.php/jhi-sbis/article/view/226>
- Claude, S., Webb, G. I. 2011. "Encyclopedia of Machine Learning".
- Cover, T. M., Hart, P. E. 1967. "Nearest Neighbor Pattern Classification". IEEE Transactions on Information Theory, 13(1), 21–27.
- DRIVENDATA. 2021. "Flu Shot Learning: Predict H1N1 and Seasonal Flu Vaccines". Tarihinde adresinden erişildi <https://www.drivendata.org/competitions/66/flu-shot-learning/>
- Eyupoglu, C., Aydin, M. A., Zaim, A. H., Sertbas, A. 2018. "An efficient big data anonymization algorithm based on chaos and perturbation techniques". Entropy, 20(5), 1–18.
- Flores, V., Keith, B. 2019. "Gradient boosted trees predictive models for surface roughness in high-speed milling in the steel and aluminum metalworking industry". Complexity, 2019.

- Gao, S., Wu, W., Lee, C., Chua, T. 2003. "A Maximal Figure-of-Merit Learning Approach", 1–8. Tarihinde adresinden erişildi papers2://publication/uuid/AA52817B-03A7-4596-AF0A-B39C2490AFB1
- Grigorios, T., Ioannis, K. 2007. "Multi-Label Classification : An Overview", 3(3), 1–13.
- Mukherjee, S., Sharma, N. 2012. "Intrusion Detection using Naive Bayes Classifier with Feature Reduction". *Procedia Technology*, 4, 119–128.
- Oprea, C., Ti, P. Ş. 2014. "Performance Evaluation of the Data Mining Classification Methods". *Analele Universităţii Constantin Brâncuşi din Târgu Jiu : Seria Economie, I*(Special number-Information society and sustainable development), 249–253.
- Öztemel, E. 2012. "*Endüstri Mühendisliğine Giriş*".
- Read, J., Pfahringer, B., Holmes, G., Frank, E. 2009. "Classifier Chains for Multi-label Classification - Machine Learning and Knowledge Discovery in Databases - Lecture Notes in Computer Science". *Machine Learning and Knowledge Discovery in Databases*, 5782, 254-269–269. Tarihinde adresinden erişildi http://link.springer.com/10.1007/978-3-642-04174-7_17%0Ahttp://www.springerlink.com/content/y70208vk20350763/
- Tan, P.-N., Steinbach, M., Kumar, V. 2014. "*Introduction to Data Mining*".
- Vuk, M., Curk, T. 2006. "ROC curve, lift chart and calibration plot". *Advances in Methodology and Statistics*, 3(1), 89–108.
- Weinberger, K. Q., Saul, L. K. 2006. "Unsupervised learning of image manifolds by semidefinite programming". *International Journal of Computer Vision*, 70(1), 77–90.
- Younes, Z., Abdallah, F., Denoeux, T., Snoussi, H. 2011. "A dependent multilabel classification method derived from the k-nearest neighbor rule". *Eurasip Journal on Advances in Signal Processing*, 2011.

EKLER

EK-1: Veri Seti Özellikleri

EK-2: BR Tahminlerinin Karmaşıklık Matrisleri

EK-3: CC Tahminlerinin Karmaşıklık Matrisleri



Ek-1:Veri Seti Özellikleri.

Öznitelik Adı	Öznitelik Türü	Öznitelik Tanımı	Eksik Veri Miktarı
respondent_id	Tam sayı	Ankete katılan kişileri temsil eden benzersiz ve rassal tekil tanımlayıcı	0
h1n1_concern	Nominal	H1N1 gribiyle ilgili endişe düzeyi (0 = Hiç endişeli, değil; 1 = Çok endişeli değil; 2 = Biraz ilgili; 3 = Çok endişeli)	92
h1n1_knowledge	Nominal	H1N1 gribiyle ilgili bilgi düzeyi (0 = Bilgisiz; 1 = Biraz bilgili; 2 = Çok bilgili)	116
behavioral_antiviral_meds	Binominal	Antiviral ilaç kullanmıştır (0 = Hayır; 1 = Evet)	71
behavioral_avoidance	Binominal	Grip benzeri semptomları olan başka kişilerle yakın temastan kaçınmıştır (0 = Hayır; 1 = Evet)	208
behavioral_face_mask	Binominal	Bir yüz maskesi satın almıştır (0 = Hayır; 1 = Evet)	19
behavioral_wash_hands	Binominal	Sık sık ellerini yıkamıştır veya el dezenfektanı kullanmıştır (0 = Hayır; 1 = Evet)	42
behavioral_large_gatherings	Binominal	Büyük buluşmaları azaltmıştır (0 = Hayır; 1 = Evet)	87
behavioral_outside_home	Binominal	Kendi evinin dışındaki insanlarla teması azaltmıştır (0 = Hayır; 1 = Evet)	82
behavioral_touch_face	Binominal	Gözlerine, burnuna veya ağzına dokunmaktan kaçınmıştır (0 = Hayır; 1 = Evet)	128
doctor_recc_h1n1	Binominal	Doktor tarafından H1N1 grip aşısı önerilmiştir (0 = Hayır; 1 = Evet)	2160
doctor_recc_seasonal	Binominal	Doktor tarafından mevsimsel grip aşısı önerilmiştir (0 = Hayır; 1 = Evet)	2160

Ek-1 - devam.

Öznitelik Adı	Öznitelik Türü	Öznitelik Tanımı	Eksik Veri Miktarı
chronic_med_condition	Binominal	Aşağıdaki kronik tıbbi rahatsızlıklardan herhangi birine sahiptir: Astım veya başka bir akciğer rahatsızlığı, diyabet, kalp rahatsızlığı, böbrek rahatsızlığı, orak hücreli anemi veya diğer anemi, nörolojik veya nöromusküler rahatsızlık, karaciğer rahatsızlığı veya kronik bir hastalık veya kronik bir hastalık için alınan ilaçlar nedeniyle zayıflamış bir bağışıklık sistemi (0 = Hayır; 1 = Evet)	971
child_under_6_months	Binominal	Altı aylıktan küçük bir çocukla düzenli yakın temas halindedir (0 = Hayır; 1 = Evet)	820
health_worker	Binominal	Sağlık çalışanıdır (0 = Hayır; 1 = Evet)	804
health_insurance	Binominal	Sağlık sigortası vardır (0 = Hayır; 1 = Evet)	12274
opinion_h1n1_vacc_effective	Nominal	Katılımcının H1N1 grip aşısının etkinliği hakkındaki görüşü (1 = Hiç etkili değil; 2 = Çok etkili değil; 3 = Bilmiyorum; 4 = Biraz etkili; 5 = Çok etkili)	391
opinion_h1n1_risk	Nominal	Katılımcının aşısız H1N1 gribine yakalanma riski hakkındaki görüşü (1 = Çok Düşük; 2 = Biraz düşük; 3 = Bilmiyorum; 4 = Biraz yüksek; 5 = Çok yüksek)	388
opinion_h1n1_sick_from_vacc	Nominal	Katılımcının H1N1 grip aşısından olunca hastalanma endişesi (1 = Hiç endişeli değil; 2 = Çok endişeli değil; 3 = Bilmiyorum; 4 = Biraz endişeli; 5 = Çok endişeli)	395
opinion_seas_vacc_effective	Nominal	Katılımcının mevsimsel grip aşısının etkinliği hakkındaki görüşü (1 = Hiç etkili değil; 2 = Çok etkili değil; 3 = Bilmiyorum; 4 = Biraz etkili; 5 = Çok etkili)	462

Ek-1 - devam.

Öznitelik Adı	Öznitelik Türü	Öznitelik Tanımı	Eksik Veri Miktarı
opinion_seas_risk	Nominal	Katılımcının aşısız mevsimsel grip hastalığına yakalanma riski hakkındaki görüşü (1 = Çok Düşük; 2 = Biraz düşük; 3 = Bilmiyorum; 4 = Biraz yüksek; 5 = Çok yüksek)	514
opinion_seas_sick_from_vacc	Nominal	Katılımcının mevsimsel grip aşısından olunca hastalanma endişesi (1 = Hiç endişeli değil; 2 = Çok endişeli değil; 3 = Bilmiyorum; 4 = Biraz endişeli; 5 = Çok endişeli)	537
age_group	Nominal	Katılımcının yaş grubu (1 = 18-34, 2 = 35-44, 3 = 45-54, 4 = 55-64, 5 = 65+)	0
education	Nominal	Beyan edilen eğitim seviyesi (<12 Years, 12 Years, Collage Graduate, Some Collage)	1407
race	Nominal	Katılımcının ırkı (White, Black, Hispanic, Other or Multiple)	0
sex	Binominal	Katılımcının cinsiyeti (Female, Male)	0
income_poverty	Nominal	2008 Nüfus Sayımı yoksulluk eşiklerine göre katılımcının hane yıllık geliri (<= \$75,000-Above Poverty, Below Poverty, > \$75,000)	4423
marital_status	Binominal	Katılımcının medeni durumu (Married, Not Married)	1408
rent_or_own	Binominal	Katılımcının barınma durumu (Rent, Own)	2042
employment_status	Nominal	Katılımcının çalışma durumu (Not in Labor Force, Employed, Unemployed)	1463
hhs_geo_region	Nominal	Katılımcının ABD Sağlık ve İnsan Hizmetleri Departmanı tarafından tanımlanan 10 bölgeyi bir coğrafi sınıflandırmayı kullanan ikametgahı (Değerler, kısa rassal karakter dizileri olarak temsil edilir)	0

Ek-1 - devam.

Öznitelik Adı	Öznitelik Türü	Öznitelik Tanımı	Eksik Veri Miktarı
census_msa	Nominal	Katılımcının ABD Nüfus Sayımı tarafından tanımlanan metropol istatistik alanlarında (MSA) ikametgahı (MSA-Not Principle City, MSA-Principle City, Non-MSA)	0
household_adults	Nominal	Hanedeki diğer yetişkinlerin sayısı, en üstte 3 olarak kodlanmıştır (0,1,2,3)	249
household_children	Nominal	Hanedeki çocuk sayısı, en üstte 3 olarak kodlanmıştır (0,1,2,3)	249
employment_industry	Nominal	Katılımcı hangi sektör tipinde istihdam edilmektedir (Değerler, kısa rassal karakter dizileri olarak temsil edilir)	13330
employment_occupation	Nominal	Katılımcının meslek türü (Değerler, kısa rassal karakter dizileri olarak temsil edilir)	13470
h1n1_vaccine	Binominal	Katılımcı H1N1 grip aşısı olmuştur (0 = Hayır; 1 = Evet)	0
seasonal_vaccine	Binominal	Katılımcı mevsimsel grip aşısı olmuştur (0 = Hayır; 1 = Evet)	0

Ek-2: BR Tahminlerinin Karmaşıklık Matrisleri

H1N1 Grip Aşısı Tahmini İçin Karmaşıklık Matrisleri

Algoritma: DET		Gerçek Sonuçlar	
		Pozitif (1)	Negatif (0)
Tahim Edilen	Pozitif (1)	5755	890
Sonuçlar	Negatif (0)	555	812

Algoritma: NAB		Gerçek Sonuçlar	
		Pozitif (1)	Negatif (0)
Tahim Edilen	Pozitif (1)	5237	677
Sonuçlar	Negatif (0)	1073	1025

Algoritma: KNN		Gerçek Sonuçlar	
		Pozitif (1)	Negatif (0)
Tahim Edilen	Pozitif (1)	5953	1229
Sonuçlar	Negatif (0)	357	473

Algoritma: GBT		Gerçek Sonuçlar	
		Pozitif (1)	Negatif (0)
Tahim Edilen	Pozitif (1)	5352	596
Sonuçlar	Negatif (0)	958	1106

Algoritma: RAF		Gerçek Sonuçlar	
		Pozitif (1)	Negatif (0)
Tahim Edilen	Pozitif (1)	6144	1199
Sonuçlar	Negatif (0)	166	503

Algoritma: DEL		Gerçek Sonuçlar	
		Pozitif (1)	Negatif (0)
Tahim Edilen	Pozitif (1)	5574	702
Sonuçlar	Negatif (0)	736	1000

Algoritma: LOG		Gerçek Sonuçlar	
		Pozitif (1)	Negatif (0)
Tahim Edilen	Pozitif (1)	5970	944
Sonuçlar	Negatif (0)	340	758

Algoritma: GLM		Gerçek Sonuçlar	
		Pozitif (1)	Negatif (0)
Tahim Edilen	Pozitif (1)	6001	996
Sonuçlar	Negatif (0)	309	706

Mevsimsel Grip Aşısı Tahmini İçin Karmaşıklık Matrisleri

Algoritma: DET		Gerçek Sonuçlar	
		Pozitif (1)	Negatif (0)
Tahim Edilen	Pozitif (1)	3292	939
Sonuçlar	Negatif (0)	990	2791

Algoritma: NAB		Gerçek Sonuçlar	
		Pozitif (1)	Negatif (0)
Tahim Edilen	Pozitif (1)	3064	930
Sonuçlar	Negatif (0)	1218	2800

Algoritma: KNN		Gerçek Sonuçlar	
		Pozitif (1)	Negatif (0)
Tahim Edilen	Pozitif (1)	3184	1229
Sonuçlar	Negatif (0)	1098	2501

Algoritma: GBT		Gerçek Sonuçlar	
		Pozitif (1)	Negatif (0)
Tahim Edilen	Pozitif (1)	3251	762
Sonuçlar	Negatif (0)	1031	2968

Algoritma: RAF		Gerçek Sonuçlar	
		Pozitif (1)	Negatif (0)
Tahim Edilen	Pozitif (1)	3529	968
Sonuçlar	Negatif (0)	753	2762

Algoritma: DEL		Gerçek Sonuçlar	
		Pozitif (1)	Negatif (0)
Tahim Edilen	Pozitif (1)	3038	682
Sonuçlar	Negatif (0)	1244	3048

Algoritma: LOG		Gerçek Sonuçlar	
		Pozitif (1)	Negatif (0)
Tahim Edilen	Pozitif (1)	3470	939
Sonuçlar	Negatif (0)	812	2791

Algoritma: GLM		Gerçek Sonuçlar	
		Pozitif (1)	Negatif (0)
Tahim Edilen	Pozitif (1)	3505	944
Sonuçlar	Negatif (0)	777	2786

Ek-3: CC Tahminlerinin Karmaşıklık Matrisleri

H1N1 Grip Aşısı Tahmini İçin Karmaşıklık Matrisleri

Algoritma: DET		Gerçek Sonuçlar	
		Pozitif (1)	Negatif (0)
Tahim Edilen	Pozitif (1)	5905	958
Sonuçlar	Negatif (0)	405	744

Algoritma: NAB		Gerçek Sonuçlar	
		Pozitif (1)	Negatif (0)
Tahim Edilen	Pozitif (1)	5109	629
Sonuçlar	Negatif (0)	1201	1073

Algoritma: KNN		Gerçek Sonuçlar	
		Pozitif (1)	Negatif (0)
Tahim Edilen	Pozitif (1)	5913	1581
Sonuçlar	Negatif (0)	397	121

Algoritma: GBT		Gerçek Sonuçlar	
		Pozitif (1)	Negatif (0)
Tahim Edilen	Pozitif (1)	5363	591
Sonuçlar	Negatif (0)	947	1111

Algoritma: RAF		Gerçek Sonuçlar	
		Pozitif (1)	Negatif (0)
Tahim Edilen	Pozitif (1)	6170	1213
Sonuçlar	Negatif (0)	140	489

Algoritma: DEL		Gerçek Sonuçlar	
		Pozitif (1)	Negatif (0)
Tahim Edilen	Pozitif (1)	5555	706
Sonuçlar	Negatif (0)	755	996

Algoritma: LOG		Gerçek Sonuçlar	
		Pozitif (1)	Negatif (0)
Tahim Edilen	Pozitif (1)	5968	931
Sonuçlar	Negatif (0)	342	771

Algoritma: GLM		Gerçek Sonuçlar	
		Pozitif (1)	Negatif (0)
Tahim Edilen	Pozitif (1)	5992	973
Sonuçlar	Negatif (0)	318	729

Mevsimsel Grip Aşısı Tahmini İçin Karmaşıklık Matrisleri

Algoritma: DET		Gerçek Sonuçlar	
		Pozitif (1)	Negatif (0)
Tahim Edilen	Pozitif (1)	3480	1065
Sonuçlar	Negatif (0)	802	2665

Algoritma: NAB		Gerçek Sonuçlar	
		Pozitif (1)	Negatif (0)
Tahim Edilen	Pozitif (1)	3085	942
Sonuçlar	Negatif (0)	1197	2788

Algoritma: KNN		Gerçek Sonuçlar	
		Pozitif (1)	Negatif (0)
Tahim Edilen	Pozitif (1)	2391	2061
Sonuçlar	Negatif (0)	1891	1669

Algoritma: GBT		Gerçek Sonuçlar	
		Pozitif (1)	Negatif (0)
Tahim Edilen	Pozitif (1)	3234	750
Sonuçlar	Negatif (0)	1048	2980

Algoritma: RAF		Gerçek Sonuçlar	
		Pozitif (1)	Negatif (0)
Tahim Edilen	Pozitif (1)	3522	990
Sonuçlar	Negatif (0)	760	2740

Algoritma: DEL		Gerçek Sonuçlar	
		Pozitif (1)	Negatif (0)
Tahim Edilen	Pozitif (1)	3146	732
Sonuçlar	Negatif (0)	1136	2998

Algoritma: LOG		Gerçek Sonuçlar	
		Pozitif (1)	Negatif (0)
Tahim Edilen	Pozitif (1)	3473	941
Sonuçlar	Negatif (0)	809	2789

Algoritma: GLM		Gerçek Sonuçlar	
		Pozitif (1)	Negatif (0)
Tahim Edilen	Pozitif (1)	3509	948
Sonuçlar	Negatif (0)	773	2782

ÖZGEÇMİŞ

Kişisel Bilgiler

Soyadı, adı : ÇINAR, Timuçin
Uyruğu : T.C.

Eğitim

Derece

Üniversite ve Bölüm

Mezuniyet tarihi

Yüksek lisans-

-

Lisans

Uludağ Üniversitesi
Endüstri Mühendisliği

2017

Yabancı Dil

İngilizce