T.C.

YILDIZ TEKNİK ÜNİVERSİTESİ

FEN BİLİMLERİ ENSTİTÜSÜ

ÇOCUK HASTALARIN TANISININ BELİRLENMESİNDE HASTA VERİLERİNDEKİ ÖZNİTELİKLERİN ÇIKARILMASI

Gamze ALTUNSAÇAN

YÜKSEK LİSANS TEZİ

Matematik Mühendisliği Anabilim Dalı

Matematik Mühendisliği Programı

Danışman

Doç. Dr. Birol ASLANYÜREK

Eş Danışman

Doç. Dr. Emrah AYDIN

Temmuz, 2024

T.C.

YILDIZ TEKNİK ÜNİVERSİTESİ

FEN BİLİMLERİ ENSTİTÜSÜ

ÇOCUK HASTALARIN TANISININ BELİRLENMESİNDE HASTA VERİLERİNDEKİ ÖZNİTELİKLERİN ÇIKARILMASI

Gamze ALTUNSAÇAN tarafından hazırlanan tez çalışması 17.07.2024 tarihinde aşağıdaki jüri tarafından Yıldız Teknik Üniversitesi Fen Bilimleri Enstitüsü Matematik Mühendisliği Anabilim Dalı, Matematik Mühendisliği Programı **YÜKSEK LİSANS TEZİ** olarak kabul edilmiştir.

Doç. Dr. Birol ASLANYÜREK Yıldız Teknik Üniversitesi Danışman	Doç. Dr. Emrah AYDIN Namık Kemal Üniversitesi Eş-Danışman

Jüri Üyeleri

Doç. Dr. Birol ASLANYÜREK, Danışman
Yıldız Teknik Üniversitesi _____

Prof. Dr. Hülya ŞAHİNTÜRK, Üye
Yıldız Teknik Üniversitesi _____

Dr. Öğr. Üyesi Mustafa Zahid GÜRBÜZ, Üye
Doğuş Üniversitesi _____

Danışmanım Doç. Dr. Birol ASLANYÜREK ve eş danışmanım Doç. Dr. Emrah AYDIN sorumluluğunda tarafımca hazırlanan "Çocuk Hastaların Tanısının Belirlenmesinde Hasta Verilerindeki Özniteliklerin Çıkarılması" başlıklı çalışmada veri toplama ve veri kullanımında gerekli yasal izinleri aldığımı, diğer kaynaklardan aldığım bilgileri ana metin ve referanslarda eksiksiz gösterdiğimi, araştırma verilerine ve sonuçlarına ilişkin çarpıtma ve/veya sahtecilik yapmadığımı, çalışmam süresince bilimsel araştırma ve etik ilkelerine uygun davrandığımı beyan ederim. Beyanımın aksinin ispatı halinde her türlü yasal sonucu kabul ederim.

Gamze ALTUNSAÇAN

İmza

Bu çalışma, “Yıldız Teknik Üniversitesi Bilimsel Araştırma Proje Koordinatörlüğü’nün TYL-2023-5673” numaralı projesi ile desteklenmiştir.

Aileme

eşime

ve

kızıma

TEŞEKKÜR

Çalışmanın yürütülmesi sırasında her zaman sabrı ve desteğiyle yanımda olup bana yol gösteren Doç. Dr. Birol ASLANYÜREK'e ve eş danışmanım Doç. Dr. Emrah AYDIN'a en içten duygularımla teşekkür ederim.

Sevgisini ve desteğini her zaman yanımda hissettiren sevgili anne ve babama, her zaman yanımda olup her konuda bana destek olan biricik eşime ve canım kızıma sonsuz sevgi ve teşekkürlerimi sunarım.

Gamze ALTUNSAÇAN

İÇİNDEKİLER

KISALTMA LİSTESİ viii
ŞEKİL LİSTESİ ix
TABLO LİSTESİ x
ÖZET xi
ABSTRACT xiii
1 GİRİŞ 1
1.1 Literatür Özeti 1
1.2 Tezin Amacı 2
1.3 Orijinal Katkı 2
2 MAKİNA ÖĞRENMESİ YAKLAŞIMLARI 4
2.1 Metin Vektörleştirme Yöntemleri 4
2.1.1 Kelime Çantası 4
2.1.2 Terim Frekansı – Ters Terim Frekansı 5
2.1.3 Kelime Vektörü (Word2Vec) 5
2.1.4 Önceden Eğitilmiş Kelime Vektörü (FastText) 6
2.1.5 Kategorik Değişkenlerin İkili Gösterim Vektörü 7
2.2 Sınıflandırma Yöntemleri 7
2.2.1 Naive Bayes 7
2.2.2 Destek Vektör Makineleri 8
2.2.3 Lineer Diskriminant Analizi 8
2.2.4 K-En Yakın Komşu Algoritması 9
2.2.5 Doğrusal Regresyon Algoritması 10
2.2.6 Lojistik Regresyon Algoritması 10
2.2.7 Rastgele Orman 10
2.2.8 Karar Ağaçları 11
3 ÇOCUK HASTALIKLARIN TANISININ BELİRLENMESİNDE HASTA VERİLERİNDEKİ ÖZNİTELİKLERİN ÇIKARILMASI 13

3.1 Verilerin Etiketlenmesi 14
3.2 Verilerin Ön İşlemesi 16
3.2.1 Wagner Fisher Algoritması ile Uzaklıkların Hesaplanması 17
3.2.2 GTSL'de Bulunmayan Sözcüklerin Tespit Edilmesi ve Listeye Eklenmesi 18
3.3 Etiketlenen Veri Setinin Betimlenmesi 19
3.4 Makine Öğrenmesi Modellerinin Eğitilmesi ve Test Edilmesi 23
3.4.1 Doğruluk Ölçütleri ve Makine Öğrenmesi Modelleri için Bir Başarı Kriterinin Tanımlanması 24
3.4.2 Veri Setinde Bulunan Normalize Edilmiş Metnin Vektörleştirilmesi 26
3.4.3 Makine Öğrenmesi Modellerinin Eğitilmesi ve Test Edilmesi 28
3.4.3.1 Kelime Çantası Yöntemi Kullanılarak Modellerin Eğitilmesi ve Test Edilmesi 28
4 SONUÇ 38
KAYNAKÇA 40
TEZDEN ÜRETİLMİŞ YAYINLAR 42

KISALTMA LİSTESİ

AUC	Area Under The Curve
BOW	Bag Of Words
DDİ	Doğal Dil İşleme
DVM	Destek Vektör Makineleri
GTSL	Geniş Türkçe Terimler Sözcük Listesi
KNN	K-Nearest Neighbors
LDA	Lineer Diskriminant Analizi
LR	Lojistik Regresyon
MNB	Multinominal Naive Bayes
OHE	One Hot Encoding
RO	Rastgele Orman
TF-IDF	Terim Frekansı – Ters Terim Frekansı
ID3	Iterative Dichotomiser 3

ŞEKİL LİSTESİ

Şekil 2.1 CBoW ve Skip-Gram Yöntemleri 6
Şekil 2.2 OHE Yöntemi ile Sınıflandırma 7
Şekil 2.3 LDA Yöntemi ile Sınıflandırma 7
Şekil 2.4 KNN Yöntemi ile Sınıflandırma 7
Şekil 3.1 Hasta Veri Tabanından Alınan Epikriz Dosya Görüntüsü 13
Şekil 3.2 Gövde Bulma Dahil Ön İşleme Yapılan Kayıtların Listesi 19
Şekil 3.3 Semptom Setinin Yer Aldığı Küme 20
Şekil 3.4 Tüm Veri Seti için Kelime Bulutu 20
Şekil 3.5 Semptom İçermeyen Grup için Kelime Bulutu 21
Şekil 3.6 Semptom İçeren Grup için Kelime Bulutu 21
Şekil 3.7 Kitle Semptomuna Sahip Veri Seti için Kelime Bulutu 22
Şekil 3.8 Şişlik Semptomuna Sahip Veri Seti için Kelime Bulutu 22
Şekil 3.9 Fıtık Semptomuna Sahip Veri Seti için Kelime Bulutu 22
Şekil 3.10 Ateş Semptomuna Sahip Veri Seti için Kelime Bulutu 23
Şekil 3.11 Ağrı Semptomuna Sahip Veri Seti için Kelime Bulutu 23
Şekil 3.12 MRMR ve Maliyet Matrisi Kullanıldığında 'Bagged Trees' Sınıflandırıcı için Karışıklık Matrisi ve En Önemli 15 Tahminci 32
Şekil 3.13 MRMR Kullanıldığında 'Fine Tree' Sınıflandırıcı için Karışıklık Matrisi ve En Önemli 15 Tahminci 33
Şekil 3.14 MRMR kullanıldığında 'Fine Tree' sınıflandırıcı için üretilen model için her semptoma karşı düşen karışıklık matrisleri 34

TABLO LİSTESİ

Tablo 3.1 Veri setinde bulunan sütunlar 14
Tablo 3.2 Etiketlenmesi planlanan genel semptomlar 14
Tablo 3.3 Etiketlenmesi planlanan detaylı semptomların bir örneği 15
Tablo 3.4 Etiketleme örneği 16
Tablo 3.5 Uzaklıkları 0'dan büyük çıkan sözcük örnekleri 18
Tablo 3.6 Semptomların geçtiği kayıt sayısı 19
Tablo 3.7 Karışıklık Matrisi 24
Tablo 3.8 Test veri seti için semptomların bulunduğu kayıt sayıları 25
Tablo 3.9 Test verisi için karışıklık matrisi 26
Tablo 3.10 Kelime arama yaklaşımına ait doğruluk ölçütleri 26
Tablo 3.11 Çoklu-Etiketleme için örnek bir kayıt 27
Tablo 3.12 Çoklu-Etiketlemenin tekli etiketlemeye dönüştürülmüş hali 27
Tablo 3.13 Kelime çantası modeli için başarı ölçütleri 29
Tablo 3.14 Kelime çantası modeli için başarı ölçütleri (maliyet matrisli) 30
Tablo 3.15 Kelime çantası modeli için MRMR kullanıldığındaki başarı ölçütleri 31
Tablo 3.16 Kelime çantası modeli için MRMR kullanıldığındaki başarı ölçütleri (maliyet matrisli) 31
Tablo 3.17 MRMR kullanıldığında 'Fine Tree' sınıflandırıcı için üretilen modelin test veri seti üzerindeki başarı ölçütleri 34
Tablo 3.18 TF-IDF modeli için başarı ölçütleri 35
Tablo 3.19 TF-IDF modeli için başarı ölçütleri (maliyet matrisli) 36
Tablo 3.20 TF-IDF modeli için MRMR kullanıldığındaki başarı ölçütleri 36
Tablo 3.21 TF-IDF modeli için MRMR kullanıldığındaki başarı ölçütleri (maliyet matrisli) 37

ÖZET

Çocuk Hastaların Tanısının Belirlenmesinde Hasta Verilerindeki Özniteliklerin Çıkarılması

Gamze ALTUNSAÇAN

Matematik Mühendisliği Anabilim Dalı
Matematik Mühendisliği Programı
Yüksek Lisans Tezi

Danışman: Doç. Dr. Birol ASLANYÜREK

Eş-Danışman: Doç. Dr. Emrah AYDIN

Günümüzde hastalıkların teşhis edilmesi, yeni tedaviler bulunması veya sağlık çalışanlarına destek olunması amacıyla çok fazla sayıda yapay zekâ temelli yöntem geliştirilmiştir. Büyük miktardaki sağlık verilerinin işlenmesi sonucu sağlık alanındaki ciddi ilerlemelere rağmen gelinen seviyenin sadece başlangıç olduğu söylenebilir.

Bir hasta, bir sağlık kurumuna başvurduğunda hastanın şikâyeti, hikayesi, test sonuçları, aldığı tanı, kullanılan ilaçlar, geçirdiği operasyonlar gibi birçok bilgi elektronik kayıt sistemlerinde tutulur. Bu bilgilerin bir kısmı yapılandırılmış haldedir ve kolayca makine öğrenmesi yöntemlerinde öznitelik olarak kullanılabilir. Semptomları da içeren klinik metinler ise yapılandırılmamış halde tutulmakta ve bu nedenle doğrudan makine öğrenmesi yaklaşımında kullanmaya uygun değillerdir. Hastalıkların tanısının konmasında semptomların çok büyük etkisi vardır. Bu nedenle metin içinden semptom çıkarımı yapılarak metin verilerinin yapılandırılmış hale dönüştürülmesi, makine öğrenmesi yaklaşımlarıyla hastalıkların tespit edilmesi ve tanı süreçlerinin iyileştirilmesi açısından önemli bir rol oynayacaktır. Bu kapsamda, bu tezin amacı çocuk hastalara ait Türkçe klinik metinlerden semptom çıkarımıdır.

Yapılandırılmamış klinik metin verilerinden semptom çıkarımı için çeşitli doğal dil işleme teknikleri kullanılmıştır. Metin verileri toplandıktan sonra öncelikle anonimleştirilmiştir. Anonimleştirme işleminin ardından çocuk hastalıkları uzmanlarının desteğiyle veri seti üzerinden semptomlar etiketlenmiştir ve ardından temizleme, dönüştürme ve normalizasyon gibi ön-işlemler yapılmıştır. Temizlenmiş verileri makine öğrenmesi yöntemlerinde öznitelik olarak kullanmak amacıyla çeşitli vektörleştirme teknikleri aracılığıyla metinler sayısal hale dönüştürülmüştür. Son olarak karar ağaçları, destek vektör makineleri, lojistik regresyon, yapay sinir ağları ve rastgele orman yöntemi gibi toplu ağaç sınıflandırıcılarıyla semptomların varlığı-yokluğu ve varsa semptomun ne olduğu tahmin edilmiştir. Özellikle karar ağacı ve toplu ağaç sınıflandırıcılarının oldukça başarılı makine öğrenmesi modelleri geliştirdiği gözlemlenmiştir.

Anahtar Kelimeler: Medikal Öznitelik Çıkarımı, Semptom Çıkarımı, Doğal Dil İşleme, Makine Öğrenmesi.

ABSTRACT

Feature Extraction From Patient Data For Diagnosing Pediatric Patients

Gamze ALTUNSAÇAN

Department of Mathematical Engineering
Master of Science Thesis

Supervisor: Assoc. Prof. Dr. Birol ASLANYÜREK

Co-supervisor: Assoc. Prof. Dr. Emrah AYDIN

Nowadays, many artificial intelligence-based methods have been developed for diagnosing diseases, discovering new treatments, or providing support. Despite significant advances in the healthcare field resulting from processing large amounts of health data, it can be said that the current level is only the beginning.

When a patient applies to a health institution, various information such as the patient's complaints, medical history, test results, diagnosis, medications, and surgeries are recorded in electronic systems. Some of this information is structured and can easily be used as features in machine learning methods. However, clinical texts that include symptoms are kept in unstructured formats and are therefore not suitable for direct use in machine learning approaches. Symptoms play a very important role in diagnosing diseases. Extracting symptoms from text converts text data into a structured format and thus plays a significant role in identifying diseases and improving diagnosis processes through machine learning approaches. In this context, the aim of this thesis is to extract symptoms from Turkish clinical texts of pediatric patients.

In this study, various natural language processing techniques were used for symptom extraction from unstructured clinical text data. After collecting the text

data, they were first anonymized. After the anonymization process, symptoms were labeled in the data set with the support of pediatric specialists, and then pre-processing steps such as cleaning, transformation and normalization were carried out. The cleaned data was converted into numerical form using various vectorization techniques to be used as features in machine learning methods. Finally, decision trees, support vector machines, logistic regression, artificial neural networks, and ensemble tree classifiers such as random forests were employed to predict the presence or absence of symptoms and, if present, to identify the specific symptoms. It was observed that especially decision tree and ensemble tree classifiers developed successful machine learning models.

Keywords: Medical Feature Extraction, Symptom Extraction, Natural Language Processing, Machine Learning.

1
GİRİŞ

1.1 Literatür Özeti

Günümüzde makine öğrenmesi yöntemleri birçok alanda olduğu gibi sağlık bilimleri alanında da sıkça kullanılmaktadır. Çalışmaların bir kısmı yapılandırılmış veriler (sayısal veya kategorik halde bulunan) üzerinden yürütülmekteyken, bir kısmı da görüntü, video, ses veya metin gibi yapılandırılmamış veriler üzerinden yürütülmektedir. Elektronik sağlık kayıtlarının tuttuğu devasa miktardaki yapılandırılmamış veri ileri klinik araştırmaların yapılması için önemli bir kaynak sağlamaktadır. Bu kapsamda metin formatındaki verilerden doğal dil işleme (DDİ) teknikleri ile özniteliklerin çıkarılması, -ki bu yaklaşım klinik konsept çıkarımı (clinical concept extraction) olarak adlandırılmaktadır [1]-, önemli araştırma alanlarından biri haline gelmiştir.

Klinik konsept çıkarımı için kullanılan yöntemler kabaca kural bazlı ve makine öğrenmesi yöntemleri şeklinde ikiye ayrılır. İlk çalışmalarda genellikle kural bazlı yöntemler kullanılmıştır. Bu alandaki ilk çalışmalardan biri kural bazlı yaklaşım kullanan Medikal Dil İşleme (Medical Language Processing) adlı projedir [2]. Doğal dil işleme topluluğu içinde kural bazlı yöntemlerde tanımlanan kuralların yetersiz kaldığı ve bu yöntemlerin makine öğrenmesi algoritmalarına göre zayıf olduğu görüşü yaygındır [3]. Bu nedenle, klinik konsept çıkarımı alanında son yıllarda makine öğrenmesi bazlı yaklaşımların kullanma sıklığı artmaktadır [1].

Elektronik sağlık kayıtlarında semptomlar; tanı, hastanın demografik bilgileri, ilaçlar gibi diğer klinik konseptlerle karşılaştırıldığında çok daha fazla yapılandırılmamış halde (serbest metin şeklinde) bulunmasına rağmen, semptom çıkarımı ile ilgili araştırma yapan çalışma sayısı çok daha azdır [4]. Oysaki klinik tanıların belirlenmesinde semptomların yapılandırılmış hale getirilmesi büyük önem arz etmektedir. Klinik metin içinde semptom çıkarımı yapan MetaMap [5] ve cTAKES [6] isimli programlar “ağrı”, “ateş” gibi genel semptomları çıkarabilmektedir. Bu programlar dışında, kural bazlı yaklaşımlar kullanarak kalp yetmezliği [7], şiddetli zihinsel hastalıklar [8], kalıcı üriner kataterleri [9] ve yapay sinir ağları kullanarak atriyal fibrilasyon hastalığı [10] gibi spesifik hastalıklara

dönük semptom çıkarımı yapan yaklaşımlar da geliştirilmiştir. Klinik metinlerden daha genel semptom çıkarma amaçlı çalışmalar için ise yapay sinir ağları temelli yaklaşımlar [4,11] geliştirilmiştir. Divita ve arkadaşları ise genel semptom çıkarımı için çeşitli makine öğrenmesi algoritmalarını karşılaştıran bir çalışma yapmışlardır [12].

Literatürde yapılan çalışmalar yetişkin semptomlarının çıkarılması üzerinedir. Ayrıca bu çalışmaların büyük kısmı İngilizce klinik metinler kullanılarak yapılmıştır ve Türkçe klinik metinler kullanılarak yapılan semptom çıkarımı çalışması bulunmamaktadır.

1.2 Tezin Amacı

Bu tezin amacı; çocuk hastalıklarının tanısının belirlenmesinde kullanmak üzere Türkçe elektronik sağlık kayıtlarındaki serbest şekilde yazılmış (yapılandırılmamış) metinler içinden hastalık semptomlarının çıkarılmasıdır.

Semptomlar hastaneye başvurma aşamasındaki başlıca nedendir ve hastalıkların teşhis edilmesinde doğrudan etkilidir. Sağlık kayıtları tutulurken hastanın hikayesi ve şikâyeti serbest metin şeklinde yazılır. Bu şekilde oluşturulan yapılandırılmamış metinler kullanarak semptomların hastalıkla ilişkisisin istatistiksel olarak analizi veya makine öğrenmesi algoritmalarıyla doğrudan hastalıkların tahmin edilmesi mümkün değildir. Bu nedenle Türkçe yapılandırılmamış klinik metinler içinden semptomların çıkarılmasına dönük geliştireceğimiz doğal dil işleme yaklaşımı çocuk hastalıklarının tanısının belirlenmesi konulu çalışmalarda kullanılabilecektir. Bu yaklaşım, ayrıca, ulusal çapta sağlık çalışanlarına destek olacak yapay zekâ temelli tanı koyan ve tedavi öneren sistemlerin geliştirilmesine katkıda bulunacaktır.

1.3 Orijinal Katkı

Literatür özetinde bahsedildiği gibi özellikle son on yılda klinik metinlerden semptom çıkarımı ile ilgili çalışmalar yoğunlaşmıştır. Fakat bu çalışmaların içinde Türkçe klinik metinlerin kullanıldığı bir yayın ya da yazılım bulunmamaktadır. Bu tez çocuk hastalıklarının tanısının belirlenmesinde kullanmak üzere Türkçe

yapılandırılmamış klinik metinler içinden semptomların çıkarılması konusunda bir örnek teşkil etmektedir.

2
MAKİNA ÖĞRENMESİ YAKLAŞIMLARI

Bu tezde makine öğrenmesi yöntemleri aracılığıyla yapılandırılmamış Türkçe metin verilerinde bulunan semptomların çıkarımı anlatılmaktadır. Bu süreçte makine öğrenmesi yöntemlerinin kullanılabilmesi için öncelikle metinlerden oluşan kayıtlar sayısal vektörlere dönüştürülecektir. Bu işlem kelime ya da metin vektörleştirme adıyla bilinir. Vektörleştirme işlemi sonunda elde edilen sayısal vektörler makine öğrenmesi modellerinde öznitelikler olarak kullanılabilecektir.

Aşağıda sırasıyla metin vektörleştirme yöntemleri ve denetimli makine öğrenmesi için sınıflandırma yöntemleri anlatılmaktadır.

2.1 Metin Vektörleştirme Yöntemleri

Kelime vektörleştirmesi, makine öğrenmesi aşamasından önce uygulanan bir yaklaşımdır. Burada metin farklı yöntemler ile sayısal veriler içeren vektörlere dönüştürülmektedir. Buradaki amaç kelimelere kendi arasında yakınlığına göre sayısal değer atamak ve kelimenin sınıf içerisinde sıklığını belirtmektir. [13]. Bu yöntem yapılandırılmamış sözel verileri makine öğrenmesi algoritmalarının işleyebileceği yapılandırılmış sayısal formatlara dönüştürür. Başlıca vektörleştirme yöntemleri aşağıda özetlenmiştir.

2.1.1 Kelime Çantası

Kelime çantası (Bag of Words - Bow) metinlerin vektörleştirilmesinde sıklıkla kullanılan yöntemlerden biridir. Bu yöntem DDİ'de metin verilerini temsil etmenin

basit ve sezgisel bir yoludur. Belgede yer alan kelimelerin varlığına ve sıklığına odaklanarak, burada yer alan kelimelerin sırasını göz ardı eder.

Bu yöntem ile her bir terim öznitelik olarak değerlendirilmektedir. Her bir özniteliğin terim içerisindeki bulunma sıklığına göre özniteliğin değeri atanarak metin sayısal bir vektöre dönüştürülmektedir.

2.1.2 Terim Frekansı – Ters Terim Frekansı

Terim Frekansı – Ters Terim Frekansı (Term Frequency - (Inverse Documet Frequency, TF-IDF), bir terimin bir belge içindeki ve bütündeki önemini değerlendiren istatistiksel bir ölçümdür. Terim frekansını (bir terimin bir belgede ne sıklıkta göründüğünü) ters belge frekansının (bir terimin belgeler arasında ne kadar benzersiz olduğu) çarpımı sonucunda TF-IDF değeri elde edilir.

Bir veri setinin metin içeren farklı belgelerden oluştuğunu varsayalım. Bu durumda bir t teriminin $d.$ belgesi içindeki terim frekansı

$$TF(t,d) = \frac{d.\text{belge içindeki } t \text{ teriminin sıklığı}}{d.\text{belge içindeki } toplam \text{ terim sayısı}} \tag{2.1}$$

formülü ile hesaplanır. IDF ise toplam belge sayısının hedef terimin geçtiği toplam belge sayısına olan oranının logaritmasıdır. Başka bir ifadeyle; bir t teriminin ters terim frekansı

$$IDF(t) = \log_{10} \frac{\text{toplam belge sayısı}}{t \text{ teriminin geçtiği toplam belge sayısı}} \tag{2.2}$$

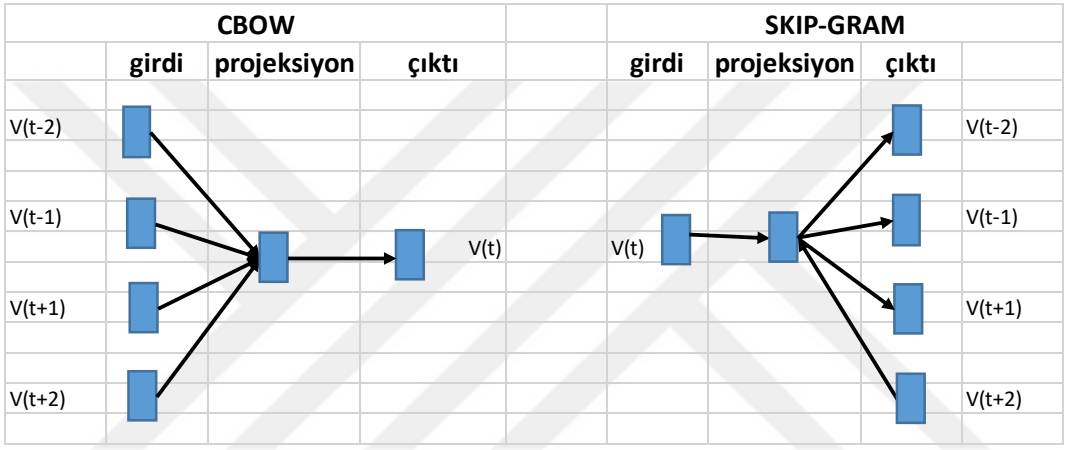
formülüyle hesaplanır. Son olarak bu iki terimin çarpımı ile bir t teriminin $d.$ belgesi içindeki TF-IDF değerine ulaşılır.

$$TF - IDF(t,d) = TF(t,d) \times IDF(t) \tag{2.3}$$

2.1.3 Kelime Vektörü (Word2Vec)

Word2Vec, Google haberlerindeki, 100 milyar kelime eğitilerek geliştirmiş bir DDİ paketidir. Kelimeler arasındaki anlamsal benzerliği, kelimelerin anlamlarını ve diğer kelimeler ile ilişkilerini yakalayarak kelimeleri vektörleştiren bir yöntemdir.

Bu vektörleştirme yönteminin CBoW ve Skip-gram adında iki ayrı tipi vardır. CBoW bir kelimenin çevresindeki kelimelerden hangisinin bu kelimenin içerisinde görülme olasılığını tahmin ederken, Skip-gram ise CBOW un aksine merkez sözcüğe verilen kelimeleri çevreleyerek çevre bağlam kelimelerini tahmin etmeye çalışır. CBoW modeli sık kullanılan kelimelerle daha iyi performans gösterirken, Skip-gram modeli seyrek kullanılan kelimelerle daha iyi performans göstermektedir. Skip-gram, CBoW modeline göre kelimeler arasındaki anlamsal ilişkileri daha iyi yakalarken, eğitim süresi daha uzundur. Şekil 2.1 de CBoW ve Skip-Gram yaklaşımları gösterilmiştir.



Şekil 2.1 CBoW ve Skip-Gram Yöntemleri

Word2Vec'in temel amacı, kelimeler arasındaki anlamsal ilişkileri, benzer kelimelerin benzer vektör temsillerine sahip olmasını sağlayacak şekilde yakalamaktır. Bu, metin sınıflandırma, duygu analizi ve makine çevirisi gibi çeşitli DDİ görevlerini gerçekleştirmeyi daha etkili hale getirir çünkü model, kelimeler arasındaki bağlamsal ve anlamsal benzerlikleri anlayabilmektedir.

2.1.4 Önceden Eğitilmiş Kelime Vektörü (FastText)

FastText 2016 yılında Facebook tarafından geliştirilen Word2Vec tabanlı açık kaynaklı bir modeldir. Bu yöntemin Word2Vec'ten farkı, kelimeleri tek varlıklar olarak temsil etmek yerine onları karakter n-gramları adı verilen daha küçük bileşenlere ayırmaktadır [14]. Kelimeleri n-gram karakter dizileri olarak temsil ederek kelime dağarcığının boyutunu etkili bir şekilde azaltabilir ve eğitim ve çıkarımın verimliliğini artırabilir. Büyük veri kümelerini hızlı ve doğru bir şekilde işlerken metin sınıflandırma ve temsil görevleri için ölçeklenebilir çözümler elde

etmektedir. Eğitimin sonunda, kelimelerin morfolojisi hakkındaki bilgileri yakalamasına ve sözlük dışında kalan kelimeleri geleneksel kelime çantası modellerine göre daha etkili bir şekilde ele almasına olanak tanımaktadır.

2.1.5 Kategorik Değişkenlerin İkili Gösterim Vektörü

Kategorik değişkenlerin ikili gösterim vektörü (One Hot Encoding - OHE) metin verilerinin içinde bulunan eşsiz kelime sayısı boyutunda vektörler oluşturan bir yöntemdir. Vektör elemanları, her bir kelimenin varlığında 1 yokluğunda 0 olacak şekilde atanır. OHE, uygulaması basit ve anlaşılması kolay bir yöntemdir. Ancak veri kümesindeki özelliklerin sayısını önemli ölçüde artırabilir ve bu durum hesaplama sorunlarına yol açabilir. Bu yöntem genelde veri çeşitliliğinin fazla olduğu ve veriler arasındaki anlamsal ilişkilerin önemli olduğu durumlarda başarısız sonuçlar üretir.

INDEKS	SEMPTOM
1	Ağrı
2	Şişlik
3	Ateş
4	Kitle
5	Fıtık
6	Kusma
7	Semptom Yok

OHE Yöntemi İle Vektörleştirilmiş Data
[0 1 0 1 0 1 0 1 1 ..... 1]
[1 1 0 1 0 1 0 1 1 ..... 0]
[0 1 0 1 0 1 0 1 1 ..... 1]
[1 1 0 1 0 1 0 1 1 ..... 1]
[0 1 0 1 0 1 0 1 1 ..... 1]
[1 1 0 1 0 1 0 2 0 ..... 0]
[0 1 0 1 0 1 0 1 1 ..... 0]

Şekil 2.2 OHE Yöntemi ile Sınıflandırma

2.2 Sınıflandırma Yöntemleri

Makine öğrenmesi modellerini eğitmek için kullanılan başlıca sınıflandırma yöntemleri aşağıda anlatılmaktadır.

2.2.1 Naive Bayes

Naive Bayes koşullu olasılık teorisini baz alan bir yöntemdir. Naive Bayes denetimli öğrenme kategorisinde yer alan, hedef değeri tahmin eden tüm özelliklerin birbirinden bağımsız olduğu varsayımıyla Bayes Teoremine dayanan bir sınıflandırma algoritmasıdır. Ayrı ayrı her sınıfın olasılığını hesaplar ve bunların içerisinden en yüksek olasılığa sahip olanı seçer. Yaygın olarak metin verileri için kullanılmasına rağmen kategorik, ikili ve sayısal özellikleri de işleyebilir. Hem ikili hem de çok sınıflı sınıflandırma için kullanılabilir. Metin sınıflandırma

problemlerinde çok popüler olarak kullanılan Naive Bayes, çok sınıflı problemlerde daha iyi sonuç vermesi ve bağımsızlık kuralı nedeniyle, diğer algoritmalara göre yüksek başarı oranına sahiptir. Naive Bayes basit ve hızlı bir algoritma olmasına rağmen özniteliklerin birbirinden bağımsız olduğunu varsayması ve sıfır olasılık (zero probability) eksik yanlarıdır. Laplace yumuşatma ile bu durumun önüne geçilmesi gerekmektedir. Aksi taktirde herhangi bir hesaplamada kullanıldığında sonuç 0 olacaktır.

Naive Bayes algoritmasının ilham aldığı Naive Bayes teoremine (2.4) te yer verilmiştir.

$$P(A|B)\ \frac{P(B|A)\,P(A)}{P(B)} \tag{2.4}$$

Burada, $P(A)$ ve $P(B)$, A ve B olaylarının gerçekleşme olasılığını temsil ederken, $P(A|B)$ B olayı verildiğinde A olayının koşullu olasılığını, $P(B|A)$ ise A olayı verildiğinde B olayının koşullu olasılığını göstermektedir.

Naive Bayes'in Multinomial Naive Bayes, Bernoulli Naive Bayes ve Gaussian Naive Bayes olarak 3 adet çeşidi bulunmaktadır.

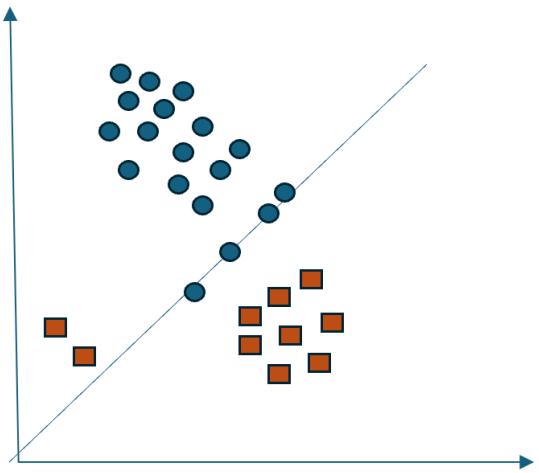
2.2.2 Destek Vektör Makineleri

Destek vektör makineleri (DVM) istatistiksel öğrenme temeline dayanan denetimli bir öğrenme algoritmasıdır. DVM, hiper düzlemler yani karar sınırları ile farklı sınıflara ait örnekleri birbirinden en uygun şekilde ayırmak için kullanılır. Yüksek boyutlu uzaylarda etkili olması, belleği verimli şekilde kullanması ve çok yönlü çekirdek fonksiyonları kullanıyor olması destek vektör makineleri algoritmasını avantajlı kılmaktadır [15]. Yüz tanıma sistemlerinden ses analizine kadar birçok alanda DVM kullanılmaktadır. Doğrusal olan ve doğrusal olmayan çeşitleri geliştirilmiştir.

2.2.3 Lineer Diskriminant Analizi

Lineer Diskriminant Analizi (LDA), denetimli makine öğreniminde çok sınıflı sınıflandırma problemlerini çözmek için kullanılan bir yaklaşımdır. Veri boyutunun

azaltılması yoluyla birden çok sınıfı birden çok özelliğe ayırır. LDA, Naive Bayes algoritmasını kullanarak ilgili veri kümesinin belirli bir çıktıya ait olup olmayacağının olasılığını hesaplar. Birden fazla nesne veya olay sınıfını birbirinden ayıran veya özelleştiren özelliklerin doğrusal bir kombinasyonunu tanımlar. Daha kolay sınıflandırabilmek için iki veya daha fazla boyutlu verileri tek boyuta indirger. LDA, ikili sınıflandırmayla sınırlı olan lojistik regresyonun aksine, çok sınıflı veri sınıflandırma problemleri için kullanılabilir. Bu nedenle LDA karar ağacı, rastgele orman veya DVM gibi diğer sınıflandırma algoritmalarının çalışmasını geliştirmek için kullanılmaktadır.

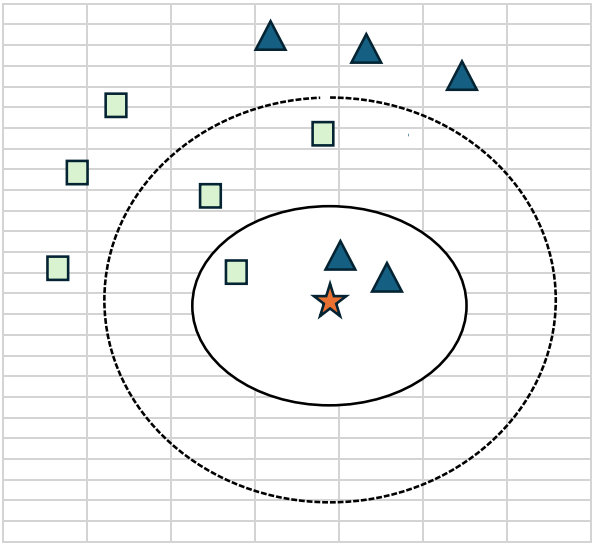


Şekil 2.3 LDA Yöntemi ile Sınıflandırma

2.2.4 K-En Yakın Komşu Algoritması

K-en yakın komşu (K-Nearest Neighbors – KNN) regresyon ve sınıflandırma için kullanılan denetimli öğrenme algoritmasıdır. Bu algoritmanın amacı, belirli bir test veri noktasına en yakın k veri noktasını bulmak ve k değerini en yakın komşuları tahmin etmek için kullanmaktır.

KNN algoritması, test noktasını, k-en yakın komşular arasında en sık görülen sınıfa atar. Başka bir deyişle komşu sayısı en fazla olan sınıf tahmin edilen sınıftır. İki veri noktası arasındaki benzerliği ölçmek için kullanılan mesafe ölçüsü, KNN algoritmasının performansını etkileyen önemli bir faktördür.



Şekil 2.4 KNN Yöntemi ile Sınıflandırma

2.2.5 Doğrusal Regresyon Algoritması

Doğrusal regresyon, değişkenler arasındaki ilişkiyi modellemek için kullanılmaktadır. Normal dağılmış iki değişken (tahmin ve sonuç değişkenleri) arasında doğrusal ilişki olup olmadığını test etmemizi olanak sağlamaktadır. Değişkenlerden biri tahmin, biri sonuç değişkenidir.

2.2.6 Lojistik Regresyon Algoritması

Lojistik regresyon (LR), bir örneğin belirli bir sınıfa ait olup olmama olasılığını tahmin etmek için kullanılan denetimli bir makine öğrenme algoritmasıdır. İki veri faktörü arasındaki ilişkiyi analiz eden istatistiksel bir algoritmadır. Girdiyi bağımsız değişkenler olarak alan ve 0 ile 1 arasında bir olasılık değeri üreten ikili sınıflandırmalar için kullanılmaktadır.

2.2.7 Rastgele Orman

Rastgele orman (RO), birden fazla karar ağacı kullanılarak geliştirilen bir yöntemdir. Kullanım kolaylığı ve esnekliği sayesinde sınıflandırma ve regresyon problemlerinde sıkça kullanılmaktadır.

RO, güçlü bir ağaç öğrenme tekniğidir. Eğitim aşamasında veri kümesi küçük parçalara ayrılarak karar ağaçları oluşturur. Ağaçları oluşturmak için CART (Classification and Regression Tree) algoritmasını kullanır. Her ağaç, veri

kümesinin rastgele bir alt kümesi kullanılarak oluşturulur. Tek tek ağaçlar arasında değişkenlik sağlayarak aşırı uyum riskini azaltır ve genel tahmin performansını artırır. Rastgele ormanlar, karmaşık verileri işleme, aşırı uyumu azaltma ve farklı ortamlarda güvenilir tahminler sağlama yetenekleriyle bilinen sınıflandırma ve regresyon fonksiyonları için yaygın olarak kullanılmaktadır.

2.2.8 Karar Ağaçları

Karar Ağaçları sınıflandırma ve regresyon problemleri için kullanılan parametrik olmayan denetimli makine öğrenmesi algoritmasıdır. Ağaç yapısına sahip olan bu algoritma kök, dallar, düğümler ve yapraklarda oluşmaktadır. Kök düğümünden başlayarak çıkan dallar karar düğümleri olarak tanımlanan iç düğümlere, iç düğümler ise tüm olası sonuçları temsil eden yaprak düğümlere doğru beslenir. Karar ağaçlarında ID3 (Iterative Dichotomiser 3), C4.5 ve CART gibi pek çok algoritma kullanılmaktadır. Karar ağaçları içerisinde bölünme noktalarını tespit edebilmek için ID3 entropi ve bilgi kazancını, C4.5 bilgi kazancı ve kazanç oranlarını, CART ise Gini yi kullanmaktadır.

2.2.8.1 Entropi

Entropi, veri kümesindeki belirsizlik miktarının ölçüsüdür. (2.5) teki formül ile hesaplanmaktadır.

$$\mathrm{H} = -\sum \mathrm{P(x)}\ \mathrm{logi}\ \mathrm{P(x)} \tag{2.5}$$

2.2.8.2 Bilgi Kazanımı

Bilgi kazanımı bir kümenin özniteliklerine göre bölünmeden öncesindeki ve sonrasındaki entropi farkının ölçüsüdür. (2.6) teki formül ile hesaplanmaktadır. S veri kümesi, D ise bölünmüş parçalardır.

$$\text{Kazanım}\ (\mathrm{S}, \mathrm{D}) = \mathrm{H(S)} - \sum_{v \in D} \frac{|V|}{|S|} H(V) \tag{2.6}$$

2.2.8.3 Gini

Gini, kümeden rastgele seçilen bir öğenin, alt kümedeki etiketlerin dağılımına göre rastgele etiketlenen bir öğenin ne sıklıkla yanlış tanımlanabileceğinin ölçüsüdür. (2.7) teki formül ile hesaplanmaktadır

$$\text{Gini (S)} = 1 - \sum_{i=1}^{c} pi^2 \qquad (2.7)$$

3
ÇOCUK HASTALIKLARIN TANISININ BELİRLENMESİNDE HASTA VERİLERİNDEKİ ÖZNİTELİKLERİN ÇIKARILMASI

Bu çalışmada, Tekirdağ Namık Kemal Üniversitesi Hastanesi'nin çocuk cerrahisi kliniğine başvuran çocuk hastaların 2010 – 2021 yılları arasındaki epikriz ve poliklinik raporları kullanılmaktadır. Hasta verilerinin bilimsel bir araştırmada kullanılması için 2021.58.02.21 protokol numaralı etik kurulu raporu alınmıştır. Epikriz raporlarında başta kimlik bilgileri olmak üzere çalışma kapsamında kullanılmayacak olan birçok özel bilgi silinerek veriler anonimleştirilmiştir. Epikriz raporlarında bulunan hikâye, şikâyet ve muayene bölümlerindeki metin kısımları öznitelik çıkarma için kullanılmıştır.

HİKAYESİ	MUAYENE	ŞİKAYETİ	TEDAVI	TANI	ICD
opere oldu	sorunsuz	kasıkta şişlik	öneri	K40 - İNGUİNAL HERNİ	K40
3 günlükken yapmış 41 haftalık doğmuş müdahalesiz kaka yapmıyo	batın ileri derecede distanmdü barsaklar palpe edilli	kaka yapamama	ADBG	Q43.1 - HİRSCHSPRUNG HAS	Q43.1
OPERE	SORUNSUZ	YUMURTASINDAN	ÖNERİ	Q53.1 - İNMEMİŞ TESTİS, UN	Q53.1
psikologa gitmişler	dilaltı bağ yok	konuşamama	konuşma terapisti KBB	K14 - DİL HASTALIKLARI	K14
sol atrofik testis nedeniyle takipli	sağ skrotumda sol skrotumüst tarafta atrofik	Yumurtasından	USG (15.2.2014) Sağ 1	Q53.1 - İNMEMİŞ TESTİS, UN	Q53.1
çocuk doktoru farketmiş	sağ testis skrotumda sol inguinal kanalda palpable	yumurtasının yerinde olmaması	öneri	Q53.1 - İNMEMİŞ TESTİS, UN	Q53.1
son 6 aydır varmış, beslenme bozuk	her 2 testis skrotumda palpable sol kordon elemanl	kaka yapamama	ADBG	K59.0 - KABIZLIK	K59.0
çiş yapamıyor diye gittiklerinde genel cerrah fıtığını farketmiş	her 2 testis skrotumda palpable sol kordon elemanl	ağrı	öneri	N48 - PENİSİN DİĞER BOZUK	N48
LUT nedeniyle takipli anne sağın da ara sıra yerinde olmadığını söyl	sol testis inguinal kanal proksimalinde palapble sağ	yumurtasından	öneri	Q53.1 - İNMEMİŞ TESTİS, UN	Q53.1
ciddi balanit atakaları	ventralde cilt kısa kordi ? mea glans ucunda	pipisinden	öneri	N48 - PENİSİN DİĞER BOZUK	N48
buji uyguluyor	perineal fistüllü anal atrezi	poposunun yerinde olmaması	öneri	Q42.3 - ANÜSÜN KONJENİTA	Q42.3
opere oldu	her 2 testis skrotumda palpable sünnetl,i sol testis ı	yumurtalarının yerinde olmaması	öneri USG	Q53.2 - İNMEMİŞ TESTİS, BİL	Q53.2
anne takipte neredeyse hiç görmediğini söyledi	sol testis ext ringde palpable sağ skrotumda pipi gör	sol yumurtasının yerinde olmaması	öneri	Q53.1 - İNMEMİŞ TESTİS, UN	Q53.1
OPERASYON PLANLANMIŞTI	SAĞ ORBİTA LATERALİNDE ŞİŞLİK, DERMOİD KİST	GÖZÜNÜN YANINDA ŞİŞLİK	ÖNERİ	I88 - LENFADENİT, NONSPESİ	I88
buji yapılıyor	perineal fistüllü anal atrezi	poposunun yerinde olmaması	öneri	Q42.2 - ANÜSÜN KONJENİTA	Q42.2
çocuk doktoru mea darlığından şüphelenmiş fucithalmic yazmıştın	işemesini gördüm çok iyi kalibrede fışkırtarak işiyor	pipisinden	öneri	N48 - PENİSİN DİĞER BOZUK	N48
1 kez açılmıştı tekrarlamış	hafif yapışıklık separe edildi	yapışıklık	öneri	Q52.5 - LABİAL FÜZYON	Q52.5
kanaması olmamaış	batın rahat, kitle palpe edilemedi anal fissür yok	kaka yapmada zorluk	öneri	K59.0 - KABIZLIK	K59.0
ağlayınca beliriyor 15 gündür farketmişler	sol kasık fıtığı sünnetsiz	kasıkta şişlik	öneri	K40 - İNGUİNAL HERNİ	K40
hidrosel için takip planlamışız	hidrosel bulgusu yok sağ testis skrotumda ama sank	pipisinden, torbasından	öneri	N48 - PENİSİN DİĞER BOZUK	N48
şu an karın ağrısı şikayeti yok	obez batın rahat ele gelen kitle yok hassasiyet yok	çocuk doktoru yönlendirilmiş	Ursofalk	K82 - SAFRA KESESİNİN DİĞE	K82
kusma yok iştahsızlık yok	sol üstte hassasiyet rebound yok	karın ağrısı	ADBG USG LFT	K76 - KARACİĞERİN DİĞER H	K76
STAR MEDİCADA SÜNNET YAPILMIŞ	GERİLEMİŞ	PİPİDE ŞİŞLİK	ÖNERİ	N48 - PENİSİN DİĞER BOZUK	N48
GÜNDÜZ İDRAR VE KULODUNU KILETİR TARZDA KAKA KAÇIRMASI V	BATIN DİSTANSİYONU YOK BARSAK SESLERİ HİPOAKT	İDRAR VE KAKA KAÇIRMA	ÖNERİ	N39.4 - ÜRİNER İNKONTİNAN	N39.4
idrar kaçırma nedeniyle tedavi görmüş	ADBG fos	kaka kaçırma	ADBG fos	K59.0 - KABIZLIK	K59.0
göbekte şişlik	umbilikal herni	göbekte şişlik	öneri	K42 - UMBİLİKAL HERNİ	K42
opere olmuştu	her 2 testis skrotumda palpable sorunsuz	yumurtasından	öneri	Q53.2 - İNMEMİŞ TESTİS, BİL	Q53.2
USG de minimal pelvikaliektazi saptanmış VCUG IVP üretrografi ist	öneri	karın ağrısı endeniyle doktora gitmişler	ADBG FOS	K59.0 - KABIZLIK	K59.0
Ekrem bey sünnet yapmış yapışıklık olmuş yeniden ameliyata alıp s	kızarık hafif şiş	pipisinden	öneri	N48 - PENİSİN DİĞER BOZUK	N48
2 yaşında sünnet olurken hidroselini 1 yıl takip edelim demişler	her 2 testis skrotumda çok belirgin hidrosel bulgusu	sol skrotal şişlik	öneri	N43.2 - HİDROSEL, DİĞER	N43.2
yapışıklık olduğunu düşünmüşler UTI yok	fimozis yok	pipisinden	öneri	N48 - PENİSİN DİĞER BOZUK	N48
midpenil hipospadias	sol testis net palpe edilemedi	pipisinden	öneri	Q54 - HİPOSPADİAS	Q54
KBBB tonsil ameliyatı yapacak	sünnetsiz	pipisinden	öneri	N48 - PENİSİN DİĞER BOZUK	N48
ucunda yapışıklık var diye düşünmüş	sünnetsiz fimozis yok	pipisinden	öneri	N48 - PENİSİN DİĞER BOZUK	N48
Mayıstan beri karnı ağrıyormuş, son 1 haftadır yine ağrısı varmış, di	barsak sesleri hipoaktif, sağ alt kadranda hassasiyet (	karın ağrısı	ADBG USG TKS TIT	R10.4 - KARIN AĞRISI DİĞER	R10.4
6 ay sonraya USG planlamışız	USG istendi	son sintigrafisi normal bulunmuş	USG istendi	R82.7 - İDRARIN MİKROBİYO	R82.7

Şekil 3.1 Hastane Veri Tabanından Alınan Epikriz Dosya Görüntüsü

Bu kısımlar ve her gelen hasta başvurusu için atanan başvuru numarası kullanılarak 40.809 satır içeren bir Excel dosyası oluşturulmuştur. Bu dosyadaki sütunların açıklaması aşağıdaki tabloda verilmiştir.

Tablo 3.1 Veri setinden bulunan sütunlar

Sütun Adı	Açıklama
Başvuru No	Kliniğe yapılan 13.603 başvuru için atanan eşsiz numara
ID	Her başvuru için hazırlanmış olan epikriz raporu hikâye, şikâyet veya muayene bölümlerini içermektedir. Bu bölümlerin her biri ayrı bir serbest metin olarak düşünüldüğünden dolayı her başvurudan 3 ayrı kayıt oluşturulmuştur ve bu kayıtlara 'ID' adında eşsiz bir numara atanmıştır.
Metin	Epikriz raporlarında bulunan hikâye, şikâyet veya muayene bölümlerinin birinden alınan metin kısmı

3.1 Verilerin Etiketlenmesi

Verilerin etiketlenmesine başlamadan önce çocuk hastalıklarında görülen temel semptomların belirlenmesi gerekmektedir. Bu semptomlar ağrı, şişlik, kitle gibi genel semptomlar olabileceği gibi karın ağrısı, boyunda şişlik, memede kitle gibi daha detaylı tanımlar içeren semptomlar da olabilir. Bunların her ikisi de önemli bilgiler içerebileceği için veriler bu semptomlarla ayrı ayrı etiketlenmiştir. Etiketleme için kullanılan semptomların listesi Tablo 3.2 ve Tablo 3.3'te verilmiştir.

Tablo 3.2 Etiketlenmesi planlanan genel semptomlar

Ağrı	İltihap	Semptom yok
Akıntı	Kan	Şişlik
Ateş	Kaşıntı	Tıkanıklık
Balgam	Kızarıklık	Travma
Çınlama	Kitle	Yanma Hissi
Fıtık	Öksürük	Yapışıklık
Halsizlik	Kusma	Döküntü
Hapşırma	Operasyon	Yanık

Tablo 3.3 Etiketlenmesi planlanan detaylı semptomların bir örneği

Ağız kokusu	Meme başında çatlak	Makatta şişlik, hemoroid
Ameliyat kesisi fıtığı	Gözde şişlik	Meme ağrısı
Anorektal malformasyon	Halsizlik yorgunluk kırgınlık, huzursuzluk	Memede kızarıklık
Anlamama	Hapşırma, aksırma, burun kaşıntısı	Meme büyüklüğü
Pelvik kitle	Hazımsızlık, yemek sonrası karın ağrısı	Anorektal malformasyon
Apse	İdrar kaçırma, yatağını ıslatma, çişini tutamama, idrarını tutamama	Bademcik iltihabı
Baş ağrısı	İdrar yaparken yanma, idrar yaparken ağrı, idrar yapmada zorluk, idrarın kokulu olması, idrarın koyu olması	Operasyon sonrası kontrol, postop kontrol
İştahsızlık	İdrar yollarında şişlik, hidroüreteronefroz	Öksürük, öksürme, hırıltılı solunum
Kan kusma	İshal, sulu kaka yapma, çok sık kaka yapma, mukuslu kaka yapma	Penis eğriliği, penis yamukluğu, kordi
Geçmeyen öksürük	Kaka kaçırma, gaita kaçırma	Erken tüylenme, kıllanma
Böbrekte şişlik, hidronefroz	Karın ağrısı (sol üst)	Pipinin küçük olması
Tırnak batması	Kasık fıtığı	Skrotumda kızarıklık, peniste kızarıklık
Dalakta büyüme, dalakta şişlik	Kasıkta şişlik, kasıkta kitle	Göğüste şişlik (rektus diyastazı)
Dışkıda kan görülmesi, kanlı dışkılama	Kaşıntı, kaşınma	Sünnet istemi

11 yıllık hasta bilgilerini içeren veri setinin çok geniş olması ve her bir kaydın etiketlenmesi için çocuk cerrahisi çalışanlarından uzman görüşlerinin alınması gerekliliği, veri setinde bulunan 40.809 kaydın tümünün etiketlenmesini oldukça zorlaştırmaktadır. Bu nedenle makine öğrenmesi modellerinin eğitimi için ilk aşamada veri setinden rastgele 1.500 kayıt seçilmiş ve etiketlenmiştir. Aşağıda rastgele seçilen 5 kaydın etiketleme örneği verilmiştir.

Tablo 3.4 Etiketleme örneği

Metin	Ana Semptom	Detaylı Semptom
antenatal tanılı hidronefroz, 4-5.ayında farkedildiğini söylüyor	şişlik	böbrekte şişlik, hidronefroz
YAKLAŞIK 2 GÜNDÜR ŞİKAYETLERİ VAR. TABLOYA ATEŞ VE SAFRASIZ KUSMA DA EKLENMİŞ. 2 GÜNDÜR HASTANEMİZ ACİL SERVİSİNE GELİYOR. DÜN ANTİBİYOTİK ANTİENFLAMATUAR VERİLMİŞ.”	Kusma Ateş	Ateş, titreme, üşüme, sıtmalanma kusma, öğürme, mide bulantısı, yediklerini çıkarma
torbasının boş olduğunu görüyorlarmış	Semptom yok	yumurtanın inmemesi, yumurtanın testiste olmaması, boş yumurtalık, gelişmemiş skrotum
TESTİSLER BİLATERAL TORBADA. SORUNSUZ.	Semptom yok	Semptom yok
“batında kistik kitle diye başvurmuşlardı Sonrasında boyutta gerileme olunca operasyonu iptal edilmiş aileye de detaylı bilgi verilmişti Kontrol USG tamamen normal her 2 adneksiyal lojda patoloji izlenmedi”	Semptom yok	Semptom yok

Zaman alıcı etiketleme sürecinin sonunda verilerin büyük kısmının semptom içermediği ve özellikle detaylı semptomlarla etiketlenen kayıt sayısının makine öğrenmesi modellerinin eğitimi açısından yetersiz kaldığı görülmüştür. Bu nedenle etiketlenen kayıt sayısına 500 kayıt daha eklenmiş ve klinik metinlerden çıkarılması planlanan semptomlar ‘ağrı’, ‘şişlik’, ‘ateş’, ‘kitle’, ‘fıtık’ ve ‘kusma’ olarak sınırlandırılmıştır. Bu semptomların seçilme sebebi etiketlenen veri setindeki en sık karşılaşılan semptomlar olmasıdır.

3.2 Verilerin Ön İşlemesi

Bu bölümde, veri setinde bulunan ‘metin’ adlı sütündeki verilerin üzerinde yapılan önişleme teknikleri aşağıda sırasıyla belirtilmiştir. Bu işlemler için Python programlama dili ve kütüphaneleri kullanılmıştır.

- Metin kısmında bulunan tüm noktalama işaretleri ve fazla boşluklar kaldırılmıştır.

- Veri setinde bulunan 40.809 kaydın 1733 adeti boş olduğundan dolayı silinmiştir.
- Metin sadece küçük harflerden oluşacak şekilde düzenlenmiştir.
- Metin kısımları sözcüklere ayrıklaştırılmıştır (word tokenization).
- Yazım yanlışlarının düzeltilmesiyle metin verilerinin normalizasyon işlemi tamamlanmıştır. Bu işlem için öncelikle tıp terimlerini de (ayrıca hekimlerin epikriz raporlarında kullandığı kısaltmaları da) içeren geniş bir Türkçe sözcük listesi (GTSL) oluşturuldu. Aşağıdaki 3 sözcük listesi kullanılarak bu geniş liste oluşturulmuştur.

    1. Türkçe doğal dil işleme için kullanılan 'Zemberek' adlı projede bulunan 1.140.838 eşsiz sözcük içeren Türkçe sözcük listesi [16]
    2. 630 sözcük içeren tıp terimleri sözlüğü [17]
    3. Veri setinde bulunan ama yukarıda bulunan listelerde yer almayan 'opere', 'USG' gibi kısaltmalar için manuel oluşturulan ayrı bir liste

  Sözü edilen GTSL'de bulunmayan ama veri setinde bulunan sözcüklerin hatalı yazıldığı varsayılarak serbest metinde bulunan tüm sözcüklerin GTSL'de bulunan her bir sözcüğe uzaklığı Wagner Fisher Algoritması ile hesaplanmıştır. Serbest metinlerde bulunan sözcüklerin uzaklıkları sıfır dışında bir değer çıktığında bu sözcüklerin hepsi incelenerek düzeltilmiş ve GTSL'ye düzeltilen sözcükler eklenmiştir.

3.2.1 Wagner Fisher Algoritması ile Uzaklıkların Hesaplanması

Wagner Fisher algoritması bir harfin eksik yazılması, fazla yazılması veya yanlış bir harf yazılması gibi durumları başarıyla dikkate alan ve sözcükler arası uzaklıkları hesaplayan bir algoritmadır. Bu algoritma kullanılarak veri setinde bulunan tüm sözcüklerin yukarıda bahsedilen GTSL'de bulunan tüm sözcüklerle olan uzaklıkları hesaplanmıştır. Veri setinde bulunan sözcük uzaklıklarının 0 çıkması durumunda bu sözcüğün yazım yanlışı içermediği kabul edilmiştir. Uzaklıkların 0 çıkmadığı durumlarda ise sözcüğün yanlış yazıldığı veya doğru yazılmasına rağmen GTSL'de olmadığı varsayılmıştır. Uzaklığı 0 olmayan toplam

4629 sözcük tespit edilmiştir. Bu sözcüklerden rastgele seçilen 5 tanesi Tablo 3.5'te verilmiştir.

Tablo 3.5 Uzaklıkları 0'dan büyük çıkan sözcük örnekleri

GTSL'de Olmayan Sözcükler	Wagner Fisher algoritmasına Göre En Yakın Sözcükler
"kbbcilere"	"['abiciler', 'abicilerde', 'abilere', 'ablilere', 'badilere', 'bakilere', 'bayilere', 'berilere', 'biciler', 'bililere', 'birilere', 'bujilere', 'cincilere', 'debilere', 'dilcilere', 'dincilere', 'ebedilere', 'edicilere', 'ekincilere', 'ekşicilere', 'emicilere', 'fobilere', 'gibilere', 'halcilere', 'hobicilere', 'hobilere', 'içicilere', 'ikicilere', 'ikincilere', 'incilere', 'iticilere', 'izcilere', 'kabilede', 'kabileme', 'kabileye', 'kabinlere', 'kabirlere', 'kabullere', 'kalecilere', 'karilere', 'keçecilere', 'keçilere', 'kedicilere', 'kedilere', 'kendilere', 'kesicilere', 'kibritlere', 'kilere', 'kililere', 'kimilere', 'kinciler', 'kirlilere', 'kirpilere', 'kişilere', 'kolilere', 'kombilere', 'komilere', 'konilere', 'köşecilere', 'kredilere', 'krokilere', 'kubbelere', 'lobicilere', 'lobilere', 'mencilere', 'mercilere', 'nebilere', 'necilere', 'rubailere', 'sabilere', 'tabiilere', 'timcilere', 'yemcilere', 'zencilere']"
"düvarı"	"duvarı"
"kawasaki"	"kafadaki', 'kaladaki', 'kanadaki', 'kapasak', 'karadaki', 'karafaki', 'kasadaki', 'kayadaki', 'kazadaki"
"hidronefloz"	"hidronefroz" "hidrofor', 'hidrosefal', 'hidrosefali"
"cağırırsınız"	"çağırırsınız"

3.2.2 GTSL'de Bulunmayan Sözcüklerin Tespit Edilmesi ve Listeye Eklenmesi

Tablo 3.5'te bulunan 'düvarı' ve 'cağırırsınız' sözcüklerinin düzeltilmiş hali olan 'duvarı' ve 'çağırırsınız' sözcükleri GTSL'de bulunduğu için bu sözcükler için yapılan düzeltme önerileri doğrudur. Fakat, 'kbb', 'kbb'ci', 'kbb'ciler', 'kbb'cilere', 'kawasaki' gibi teknik terimler GTSL'de bulunmadığı için 'kbbcilere' ve 'kawasaki' sözcükleri için verilen öneriler doğru değildir. Diğer yandan 'hidronefloz' sözcüğü için birçok öneri verilmiştir ve önerilenlerin içinde olan 'hidronefroz' aranan doğru sözcüktür. Sonuç olarak 'kbb'cilere' ve 'kawasaki' sözcükleri GTSL'ye eklenerek Wagner Fisher algoritmasının listedeki tüm sözcükler için doğru öneriler vermesi sağlanmıştır. 5 örnek için verilen bu uygulama 4629 sözcüğün tümüne uygulanmıştır.

Ham veri işlendikten sonra her bir kelime için gövde bulunmuştur.

METİN	SEMPTOM AN	SEMPTOMLA	CLEAN_METIN
YARA YERİ TEMİZ SORUNSUZ. PANSUMAN YAPıLDı. ÖNERİ.PATAOLOJİ DEĞERLENDİRİLDİ. 03/06/2020 PATOLOJİ:AKUT APANDİSİT, APPENDEKTOMİ MATERYALİ.	Semptom YOK	Semptom YOK	yara yer temiz sorun pansuman yapıldı öneri patoloji değer 03 06 2020 patoloji akut apandisit apendektomi materyal
SU FıTıĞı	Fıtık	skrotumda kitle, yumurtalıkların şiş olması, şiş yumurtalık	su fıtık
BOYUNDA ŞİŞLİK	Şişlik		boyun şiş
İDRAR TORBASıNıN ARKASıNDA LEZYON GÖRÜLMÜŞ VE TARAFıMıZA YÖNLENDİRİLMİŞ. VCUG İSTENMİŞ TRAKYA TıPTA ÇEKTİRİLEBİLECEĞİ AÇıKLANDı	Kitle		idrar torba arka lezyon görü ve taraf yön vcug iste Trakya tıp çektiri aç
akciğerlerinden	Semptom YOK	Semptom YOK	akciğer

Şekil 3.2 Gövde Bulma Dahil Ön İşleme Yapılan Kayıtların Listesi

3.3 Etiketlenen Veri Setinin Betimlenmesi

Toplam 40.809 kayıttan 2000 tanesi yukarı bahsedildiği gibi önceden etiketlenmişti. Etiketlenen veri setinin yukarıda belirtilen yaklaşımla normalize edilmesi sonucu oluşan boş metin verileri ve konuyla ilgisiz kayıtlar tespit edilmiş ve silinmiştir. Geriye 1.890 etiketli kayıt kalmıştır.

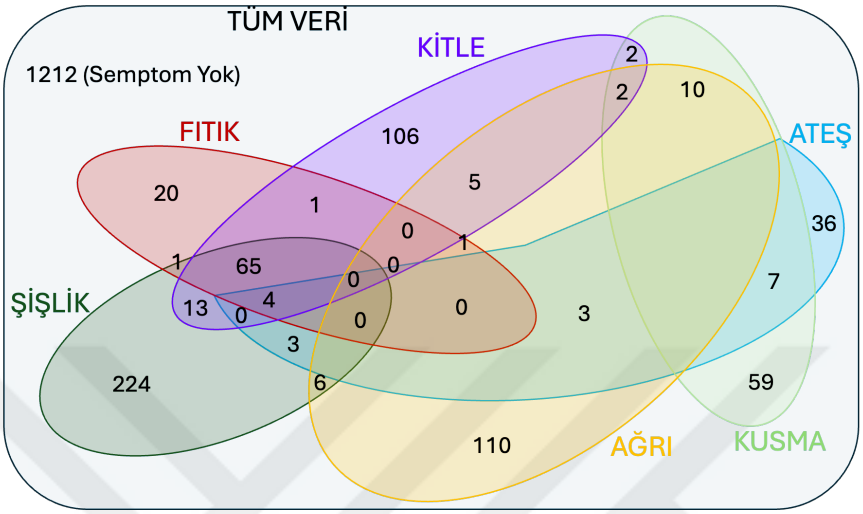
Normalize edilmiş 1890 adet veride altı adet semptom (‘ağrı’, ‘şişlik’, ‘ateş’, ‘kitle’, ‘fıtık’ ve ‘kusma’) için etiketleme bulunmaktadır. Bu verilerden 678 adet kayıtta semptom bulunurken, 1212 adet kayıtta herhangi bir semptom bulunmamaktadır. Tablo 3.6’da hangi semptom için kaç adet kayıt bulunduğu bilgisi verilmiştir.

Tablo 1.6 Semptomların geçtiği kayıt sayısı

Semptom Adı	**Kayıt Sayısı**
Ağrı	137
Şişlik	316
Kitle	198
Ateş	53
Kusma	80
Fıtık	92

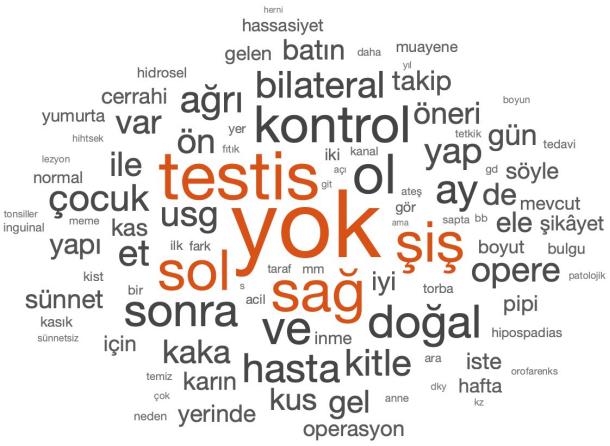
Veriler içerisinde birden fazla semptomun yer aldığı kayıtlar yer almaktadır. Örneğin; sadece ‘fıtık’ ve ‘kitle’ semptomunun aynı anda bulunduğu 1 adet kayıt

varken, fıtık, kitle, şişlik ve ateşin birlikte bulunduğu 4 adet kayıt vardır. ‘Kusma’ ile ‘fıtık’ semptomlarının veya ‘kusma’ ile ‘şişlik’ semptomlarının birlikte aynı anda bulunduğu herhangi bir kayıt bulunmamaktadır. Şekil 3.3’te tüm semptom setinin yer aldığı kümeler kesişimleriyle birlikte verilmiştir.



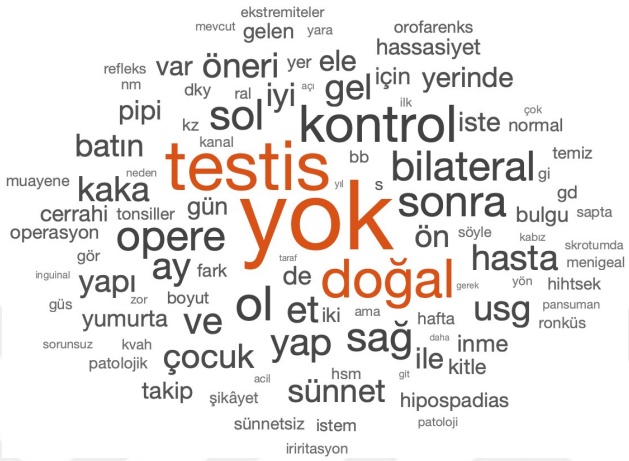
Şekil 3.3 Semptom Setinin Yer Aldığı Küme

Normalize edilmiş 1890 adet veride, bütün halinde, semptomların varlığı-yokluğu şeklinde ve altı semptom için ayrı gruplandırılarak baskın sözcükler kelime bulutları (word clouds) şeklinde üretilmiştir. Şekil 3.4’te verilen kelime bulutuna göre baskın sözcükler ‘yok’, ‘testis’, ‘şiş’, ‘sol’ ve ‘sağ’ olarak söylenebilir.



Şekil 3.4 Tüm Veri Seti için Kelime Bulutu

Şekil 3.5 ve Şekil 3.6'da ise sırasıyla semptom içermeyen ve semptom içeren gruplara ait kelime bulutları verilmiştir. Semptom içermeyen grupta 'yok' sözcüğü, semptom içeren gruba göre daha baskındır. Bunun dışında semptom içeren grupta 'şiş', 'ağrı' gibi semptomların yanında, 'sağ', 'sol' gibi sözcükler öne çıkarken, diğer grupta 'doğal' ve 'testis' sözcükleri öne çıkmaktadır.



Şekil 3.5 Semptom İçermeyen Grup için Kelime Bulutu



Şekil 3.6 Semptom İçeren Grup için Kelime Bulutu

Semptom bazında ayrı ayrı gruplara ait kelime bulutları ise Şekil 3.7 - Şekil 3.12 arasında verilmiştir. Örneğin; 'kitle' semptomu için 'kitle' sözcüğünün yanında 'sağ', 'sol', 'hidrosel' ve 'testis' sözcüklerinin baskın olduğu görülmektedir. Diğer gruplarda da semptom isimlerinin baskın sözcükler arasında olduğu gözlemlenmektedir.



Şekil 3.7 Kitle Semptomuna Sahip Veri Seti için Kelime Bulutu



Şekil 3.8 Şişlik Semptomuna Sahip Veri Seti için Kelime Bulutu



Şekil 3.9 Fıtık Semptomuna Sahip Veri Seti için Kelime Bulutu



Şekil 3.10 Ateş Semptomuna Sahip Veri Seti için Kelime Bulutu



Şekil 3.11 Ağrı Semptomuna Sahip Veri Seti için Kelime Bulutu

3.4 Makine Öğrenmesi Modellerinin Eğitilmesi ve Test Edilmesi

Etiketlenmiş ve normalize edilmiş 1.890 kaydın %90'ı eğitim, %10'u ise test için ayrılmıştır. Eğitim için ayrılan veri seti kullanılarak çeşitli makine öğrenmesi modelleri eğitilecek ve ardından eğitilen modellerin başarısı test verisi üzerinden ölçülecektir.

Aşağıda öncelikle doğruluğu ölçmek için kullanılacak ölçütler (metrikler) tanımlanmış ve ardından makine öğrenmesi modellerinin başarılı olup olmadığını

tespit etmek için uygun bir başarı kriter ortaya konmuştur. Son olarak ise bu ölçütler ve başarı kriteri kullanılarak modellerin başarısı test verisi üzerinden test edilmiştir.

3.4.1 Doğruluk Ölçütleri ve Makine Öğrenmesi Modelleri için Bir Başarı Kriterinin Tanımlanması

Sınıflandırma sonucunda modelin tahminlemesine ait doğruluk çeşitli metrikler ile ölçülmektedir. Karışıklık matrisi (Confusion Matrix), gerçek sınıf ile tahmin edilmiş sınıf arasındaki ilişkileri gösteren bir tablodur. Karışıklık matrisinde yatayda gerçek değerler, dikeyde tahmin edilen değerler gösterilmektedir. Pozitif olarak sınıflandırdığımız ve gerçekte pozitif olan değerlerin sayısı doğru pozitif (DP), pozitif olarak sınıflandırdığımız ancak gerçekte negatif olan değerlerin sayısı yanlış pozitif (YP), negatif olarak sınıflandırdığımız ve gerçekte de negatif olan değerlerin sayısı doğru negatif (DN), negatif olarak sınıflandırdığımız ancak gerçekte pozitif olan değerlerin sayısı yanlış negatif (YN) olmak üzere Tablo 3.7'de karışıklık matrisi gösterilmiştir.

Tablo 3.7 Karışıklık Matrisi

		Tahmin	
		Pozitif	Negatif
Gerçek	Pozitif	DP	YN
	Negatif	YP	DN

Modelin performansını ölçmek için karışıklık matrisinden elde edilen duyarlılık, özgüllük, kesinlik, F1 skoru, doğruluk, dengeli doğruluk ve eğri altındaki alan (Area Under the Curve - AUC) ölçütleri kullanılacaktır. AUC, alıcı işletim karakteristiği (Receiver Operating Characteristic) eğrisinin altındaki alanı ifade eder ve 1 değerine yaklaştığında yöntem ideal tahmin başarısına yaklaşırken, 0.5

değerine yaklaştığında yöntemin rastgele sınıflandırma yaptığı (başarısız olduğu) anlamı çıkar. Diğer ölçütlerin tanımları (3.1)-(3.6)'da verilmiştir.

$$\boldsymbol{Duyarlılık} = \frac{\boldsymbol{DP}}{\boldsymbol{DP + YN}} \tag{3.1}$$

$$\text{Ö}\boldsymbol{zg}\text{ü}\boldsymbol{ll}\text{ü}\boldsymbol{k} = \frac{\boldsymbol{DN}}{\boldsymbol{DN + YP}} \tag{3.2}$$

$$\boldsymbol{Kesinlik} = \frac{\boldsymbol{DP}}{\boldsymbol{DP + YN}} \tag{3.3}$$

$$\boldsymbol{F1\ skoru} = \frac{\boldsymbol{2 * (Kesinlik * Duyarlılık)}}{\boldsymbol{Kesinlik + Duyarlılık}} \tag{3.4}$$

$$\boldsymbol{Do\check{g}ruluk} = \frac{\boldsymbol{DP + DN}}{\boldsymbol{DP + YP + \ YN + DN}} \tag{3.5}$$

$$\boldsymbol{Dengeli\ Do\check{g}ruluk} = \frac{\text{Ö}\boldsymbol{zg}\text{ü}\boldsymbol{ll}\text{ü}\boldsymbol{k + Duyarlılık}}{\boldsymbol{2}} \tag{3.6}$$

Bu çalışmada kullanılan doğruluk ile ilgili ölçüler yukarıda tanımlandı. Şimdi ise sırayla test verisi üzerinden başarı kriterleri tarif edilecektir. Normalize edilmiş halde bulunan test veri setindeki kayıtlar, eğer altı semptomu temsil eden 'ağrı', 'şiş', 'kitle', 'ateş', 'fıtık' veya 'kus' sözcüklerinden birini içeriyorsa, 'semptom var' şeklinde, aksi takdirde 'semptom yok' şeklinde olacak şekilde sınıflandırıldığını varsayalım. Bu takdirde; gerçek dağılımları Tablo 3.8'de verilen semptomların karışıklık matrisi Tablo 3.9'daki şekilde olacaktır.

Tablo 3.8 Test veri seti için semptomların bulunduğu kayıt sayıları

Semptom Adı	Kayıt Sayısı
Ağrı	15
Şişlik	27
Kitle	13
Ateş	5
Kusma	7
Fıtık	9

Tablo 3.9 Test verisi için karışıklık matrisi

		Tahmin	
		Pozitif	Negatif
Gerçek	Pozitif	116	13
	Negatif	3	51

'Anahtar sözcük arama' olarak adlandırılabilecek bu yaklaşım için hesaplanmış doğruluk ölçütleri Tablo 3.10'da verilmiştir. Gerek eğitim setinde gerekse de test setinde sınıflar dengesiz olduğu için özellikle doğruluk ölçütü aldatıcı olabilir. Bu nedenle, bu tip durumlarda dengeli doğruluk, F1 skoru ve AUC gibi ölçütler daha güvenilirdir. Eğitilen makine öğrenmesi modellerine ait ölçütler başta dengeli doğruluk, F1 skoru ve AUC olmak üzere aşağıdaki değerlerin altında kalması durumunda, bu modeller başarısız sayılacaktır. 'Anahtar sözcük arama' yaklaşımı doğrudan sözcüğün geçtiği durumları pozitif olarak tahmin etmekle birlikte; semptomların ismi geçmeden semptomların daha karmaşık ifade edildiği durumlarda yanlış negatif (*YN*), semptomların isim vererek yokluğu anlatıldığı durumlarda yanlış pozitif (*YP*) sonuçlar üretecektir. Başarılı bir şekilde eğitilen makine öğrenmesi modellerinin bu hataları azaltması beklenir.

Tablo 3.10 Kelime arama yaklaşımına ait doğruluk ölçütleri

Doğruluk	0.88
Dengeli Doğruluk	0.87
Duyarlılık	0.85
Özgüllük	0.90
Kesinlik	0.80
F1 skoru	0.82

3.4.2 Veri Setinde Bulunan Normalize Edilmiş Metnin Vektörleştirilmesi

Veri setinde her bir kayıt hiçbir semptom içermeyebilir, bir semptom içerebilir veya birden fazla semptom içerebilir. Bu nedenle bu haliyle problem bir çoklu-etiketleme örneğidir (bak Tablo 3.11). Bu durumda altı semptomun her birine ayrı ayrı karşılık düşecek şekilde kayıtlar çoklanır ve veri setine '0' ve '1' kategorik değişkenlerini

alabilen yeni bir öznitelik (‘semptom no’) eklenirse problem tekli-etiket haline almış olur (bak Tablo 3.12).

Tablo 3.11 Çoklu-Etiketleme için örnek bir kayıt

ID	Normalize Veri	Semptom (çıktı)
S2296623	"gün ön kus başla sonra kus ishal devam et bugün ateş eklen"	Kusma, ateş

Tablo 3.12 Çoklu-Etiketlemenin tekli etiketlemeye dönüştürülmüş hali

ID	Normalize Veri	Semptom No	Çıktı
S2296623	"gün ön kus başla sonra kus ishal devam et bugün ateş eklen"	1 (ağrı)	0
S2296623	"gün ön kus başla sonra kus ishal devam et bugün ateş eklen"	2 (şişlik)	0
S2296623	"gün ön kus başla sonra kus ishal devam et bugün ateş eklen"	3 (ateş)	1
S2296623	"gün ön kus başla sonra kus ishal devam et bugün ateş eklen"	4 (kitle)	0
S2296623	"gün ön kus başla sonra kus ishal devam et bugün ateş eklen"	5 (fıtık)	0
S2296623	"gün ön kus başla sonra kus ishal devam et bugün ateş eklen"	6 (kusma)	1

Bu çalışmada veri seti tekli-etiketlenmiş hale dönüştürüldükten sonra eğitim veri seti 10.206, test veri seti ise 1134 örnekten oluşur hale gelmiştir. Bu işlemden sonra makine öğrenmesi modelleri üretebilmek için normalize edilmiş metin kelime çantası (BOW), TF-IDF, $\boldsymbol{N = 2}$ ve $\boldsymbol{N = 3}$ için N-gram yöntemi ve FastText yöntemleriyle ayrı ayrı sayısal vektörler haline dönüştürülmüştür. FastText yöntemi yukarıda belirtilen başarı kriterlerini aşamadığı için, N-gram yöntemleri ise hem veri setinin boyutunu ciddi şekilde artırdığı ve yine Fasttext gibi yukarıda belirtilen başarı kriterlerini aşamadığı için, sadece diğer vektörleştirme yöntemlerine ait sonuçlar bu çalışmada verilmiştir.

3.4.3 Makine Öğrenmesi Modellerinin Eğitilmesi ve Test Edilmesi

Makine öğrenmesi modellerinin eğitimi sırasında 10 katlı çapraz doğrulama tekniği kullanılmıştır. Matlab Sınıflandırma Öğreticisi (Classification Learner) adlı araç kutusu kullanılarak çeşitli makine öğrenmesi yöntemleri ile tahminlemeler yapılmıştır. Aşağıda sırasıyla farklı vektörleştirme yöntemleri için sonuçlar verilmektedir.

3.4.3.1 Kelime Çantası Yöntemi Kullanılarak Modellerin Eğitilmesi ve Test Edilmesi

Kelime çantası yöntemi uygulandığında normalize edilmiş metinler 2146 elemandan oluşan sayısal vektörlere dönüştürülmüştür. Bunlara yukarıda bahsedilen 'semptom No' adlı değişken de eklendiğinde, veri seti 2147 adet öznitelik ve 1 çıktı değişkeninden oluşmaktadır.

Tablo 3.13'de Matlab Sınıflandırma Öğreticisi adlı araç kutusunda bulunan çeşitli sınıflandırıcı yöntemlerin uygulanması sonucu elde edilen sonuçlar verilmiştir. Tablolarda her bir metrik için başarı oranları en yüksek olan değerler kırmızı ile belirtilmiştir. Doğruluk ölçütü en yüksek çıkan sınıflandırıcılar karar ağacı olan 'Fine Tree' (0,98), rastgele orman tipinde topluluk sınıflandırıcıları olan 'Boosted Trees' (0,98) ve 'Bagged Trees' (0,97) olmuştur. Dengeli doğruluk açısından en başarılı olanlara 'Fine Tree' (0,94) ve 'Bagged Trees' (0,92) yanında rastgele orman tipinde topluluk sınıflandırıcıları olan 'RUSBoosted Trees' (0,92) katılmıştır. F1 skoru açısından en başarılı olan 'Fine Tree' (0,84) olurken, AUC ölçüsü açısından 'Boosted Trees' (0,99), 'Bagged Trees' (0,97) ve 'RUSBoosted Trees' (0,97) öne çıkmaktadır. Destek vektör makinleri ('SVM'), lojistik regresyon ('Eff. Logistic Regression') ve yapay sinir ağları ('Wide Neural Network', 'Bilayered Neural Network') gibi yöntemler kullanan sınıflandırıcılar ise görece daha başarısız olmuştur.

Veri setinin dengesiz olması nedeniyle dengeli doğruluk değerlerinin görece düşük çıktığı görülmektedir. Bu sorunu telafi etmek için maliyet matrisi (cost matrix) tekniği kullanılmıştır. Maliyet matrisinde, yanlış tahmin veya doğru tahmin maliyetleri verideki dengesizlik oranı dikkate alarak belirlenir. Böylece modeller

sayısı az olan sınıfı daha fazla bulma yönünde bir eğilime girerler. Bu yaklaşım uygulandığında elde edilen sonuçlar Tablo 3.14'te verilmiştir.

Tablo 3.13 Kelime çantası modeli için başarı ölçütleri

		Doğruluk	Dengeli Doğruluk	Duyarlılık	Özgüllük	Kesinlik	F1 Skoru	AUC
Fine Tree	Validasyon	0,98	0,91	0,83	0,99	0,89	0,86	0,90
	Test	0,98	0,94	0,89	0,98	0,79	0,84	0,95
Medium Tree	Validasyon	0,96	0,80	0,61	0,99	0,87	0,71	0,80
	Test	0,96	0,82	0,64	0,98	0,78	0,71	0,82
Coarse Tree	Validasyon	0,95	0,71	0,42	1,00	0,90	0,58	0,72
	Test	0,96	0,75	0,50	0,99	0,83	0,62	0,77
SVM (linear)	Validasyon	0,91	0,50	0,02	0,99	0,15	0,03	0,73
	Test	0,93	0,50	0,00	1,00	-	-	0,62
SVM (cubic)	Validasyon	0,91	0,52	0,05	0,99	0,26	0,08	0,80
	Test	0,93	0,50	0,00	1,00	-	-	0,62
SVM (coarse Gaussian)	Validasyon	0,92	0,50	0,00	1,00	-	-	0,75
	Test	0,93	0,50	0,00	1,00	-	-	0,66
Eff. Logistic Regression	Validasyon	0,92	0,51	0,02	1,00	0,57	0,03	0,82
	Test	0,93	0,50	0,00	1,00	-	-	0,81
Boosted Trees	Validasyon	0,97	0,84	0,67	1,00	0,94	0,79	0,98
	Test	0,98	0,88	0,78	0,99	0,84	0,81	0,99
Bagged Trees	Validasyon	0,98	0,92	0,85	0,99	0,90	0,87	0,98
	Test	0,97	0,92	0,86	0,98	0,76	0,81	0,97
RUSBoosted Trees	Validasyon	0,97	0,89	0,81	0,98	0,77	0,79	0,95
	Test	0,97	0,92	0,86	0,98	0,76	0,81	0,97
Wide Neural Network	Validasyon	0,95	0,78	0,59	0,98	0,69	0,64	0,82
	Test	0,95	0,83	0,68	0,97	0,59	0,63	0,85
Bilayered Neural Net.	Validasyon	0,93	0,71	0,45	0,97	0,59	0,51	0,77
	Test	0,90	0,74	0,55	0,93	0,36	0,44	0,83
Anahtar Sözcük Arama	Test	0,88	0,87	0,85	0,90	0,80	0,82	-

Tablo 3.14 Kelime çantası modeli için başarı ölçütleri (maliyet matrisli)

		Doğruluk	Dengeli Doğruluk	Duyarlılık	Özgüllük	Kesinlik	F1 Skoru	AUC
Fine Tree	Validasyon	0,97	0,93	0,88	0,98	0,77	0,82	0,91
	Test	0,97	0,93	0,88	0,97	0,71	0,78	0,88
Medium Tree	Validasyon	0,95	0,89	0,82	0,96	0,66	0,73	0,90
	Test	0,96	0,92	0,88	0,96	0,61	0,72	0,90
Coarse Tree	Validasyon	0,66	0,72	0,79	0,65	0,16	0,27	0,78
	Test	0,62	0,71	0,82	0,61	0,13	0,22	0,74
SVM (linear)	Validasyon	0,72	0,69	0,65	0,72	0,17	0,27	0,75
	Test	0,74	0,75	0,76	0,74	0,17	0,28	0,80
SVM (cubic)	Validasyon	0,82	0,71	0,58	0,84	0,23	0,33	0,77
	Test	0,83	0,73	0,62	0,84	0,22	0,32	0,79
SVM (coarse Gaussian)	Validasyon	0,69	0,71	0,74	0,69	0,17	0,27	0,70
	Test	0,69	0,71	0,74	0,69	0,14	0,24	0,73
Eff. Logistic Regression	Validasyon	0,79	0,78	0,77	0,80	0,24	0,37	0,85
	Test	0,77	0,78	0,79	0,77	0,20	0,31	0,85
Boosted Trees	Validasyon	0,97	0,92	0,87	0,98	0,76	0,81	0,96
	Test	0,97	0,92	0,87	0,98	0,76	0,81	0,98
Bagged Trees	Validasyon	0,97	0,94	0,90	0,98	0,78	0,84	0,98
	Test	0,97	0,93	0,88	0,98	0,74	0,80	0,99
RUSBoosted Trees	Validasyon	0,97	0,90	0,82	0,98	0,77	0,79	0,95
	Test	0,97	0,94	0,91	0,98	0,74	0,82	0,97
Wide Neural Network	Validasyon	0,95	0,78	0,59	0,98	0,68	0,63	0,82
	Test	0,94	0,81	0,66	0,96	0,53	0,59	0,84
Bilayered Neural Net.	Validasyon	0,89	0,74	0,56	0,92	0,36	0,44	0,78
	Test	0,88	0,73	0,57	0,90	0,29	0,38	0,84
Anahtar Sözcük Arama	Test	0,88	0,87	0,85	0,90	0,80	0,82	-

MRMR (Minimum Redundancy Maksimum Relevance), Minimum Artıklık Maksimum Uygunluk anlamına gelen bir özellik seçim algoritmasıdır. MRMR'nin en önemli özelliği minimal-optimum özellik seçmesidir. Minimum sayıda faydalı özelliğin seçilmesi hafıza tükemi, modeli hızlandırması ve verimli olması nedenlerinden dolayı tercih edilmektedir. Bu çalışmada MRMR algoritması uygulanarak en önemli 1500 öznitelik seçilmiştir. Ardından bu yeni öznitelik kümesi için maliyet fonksiyonu kullanılmadan ve kullanılarak, yukarıda en başarılı olan 'Fine Tree', 'Medium Tree', 'Boosted Trees', 'Bagged Trees' ve 'RUSBoosted Trees' sınıflandırıcıları için modeller eğitilmiş ve test edilmiştir. Elde edilen sonuçlar; maliyet fonksiyonu olmadığı durum için Tablo 3.15, maliyet fonksiyonu olduğu durum için Tablo 3.16'da verilmiştir. MRMR algoritması ile en

önemli 1000 öznitelik seçildiğinde başarılar düşmeye başladığı için bu çalışmada ilgili sonuçlara yer verilmemiştir.

Tablo 3.15 Kelime çantası modeli için MRMR kullanıldığındaki başarı ölçütleri

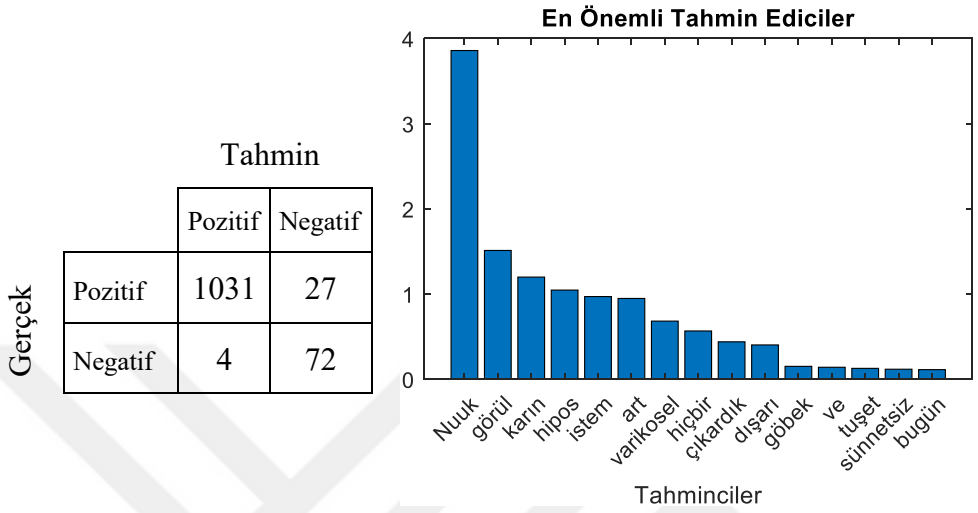
		Doğruluk	Dengeli Doğruluk	Duyarlılık	Özgüllük	Kesinlik	F1 Skoru	AUC
Fine Tree	Validasyon	0,98	0,90	0,81	0,99	0,88	0,84	0,91
	Test	0,98	0,96	0,93	0,99	0,83	0,88	0,93
Medium Tree	Validasyon	0,96	0,80	0,61	0,99	0,86	0,71	0,80
	Test	0,97	0,84	0,70	0,99	0,82	0,75	0,84
Boosted Trees	Validasyon	0,97	0,84	0,70	0,99	0,82	0,75	0,97
	Test	0,98	0,90	0,82	0,99	0,83	0,82	0,99
Bagged Trees	Validasyon	0,98	0,90	0,81	0,99	0,89	0,85	0,97
	Test	0,98	0,95	0,91	0,98	0,79	0,85	0,97
RUSBoosted Trees	Validasyon	0,97	0,88	0,77	0,98	0,79	0,78	0,95
	Test	0,97	0,92	0,87	0,98	0,76	0,81	0,97
Anahtar Sözcük Arama	Test	0,88	0,87	0,85	0,90	0,80	0,82	-

Tablo 3.16 Kelime çantası modeli için MRMR kullanıldığındaki başarı ölçütleri (maliyet matrisli)

		Doğruluk	Dengeli Doğruluk	Duyarlılık	Özgüllük	Kesinlik	F1 Skoru	AUC
Fine Tree	Validasyon	0,97	0,93	0,88	0,98	0,76	0,81	0,92
	Test	0,97	0,95	0,93	0,97	0,70	0,80	0,93
Medium Tree	Validasyon	0,92	0,88	0,83	0,92	0,48	0,61	0,90
	Test	0,96	0,92	0,89	0,96	0,61	0,72	0,90
Boosted Trees	Validasyon	0,97	0,91	0,85	0,98	0,76	0,80	0,96
	Test	0,97	0,93	0,88	0,98	0,74	0,81	0,98
Bagged Trees	Validasyon	0,97	0,93	0,89	0,98	0,76	0,82	0,97
	Test	0,97	0,96	0,95	0,97	0,73	0,82	0,98
RUSBoosted Trees	Validasyon	0,97	0,89	0,79	0,98	0,79	0,79	0,95
	Test	0,97	0,92	0,87	0,98	0,74	0,80	0,97
Anahtar Sözcük Arama	Test	0,88	0,87	0,85	0,90	0,80	0,82	-

Tablo 3.16'da görüldüğü gibi MRMR algoritması ve maliyet matrisi birlikte uygulandığında 'Bagged Trees' sınıflandırıcısının test seti üzerindeki dengeli doğruluğu 0,96'ya çıkmıştır. Bu modelin karışıklık matrisi ve tahminleme aşamasında kullandığı en önemli 15 tahminci aşağıdaki şekilde gösterilmektedir. Bu model semptom isimlerini doğrudan içermeyen, en önemli tahmincisi 'Nuuk'

olan tahmincileri kullanmıştır. Bu tahmincilerin seçilmesi açık ve beklenen bir durum değildir. 'Nuuk' tahmincisi normalize edilmemiş veri setinde bulunan 'Nuck kanalında' otomatik çevrilmiştir ve aslında 'Nuck' sözcüğünü temsil etmektedir.



Şekil 3.12 MRMR ve maliyet matrisi kullanıldığında 'Bagged Trees' sınıflandırıcı için karışıklık matrisi ve en önemli 15 tahminci

Tablo 3.15'te ise görüldüğü gibi MRMR algoritmasıyla birlikte 'Fine Tree' sınıflandırıcısının test seti üzerindeki dengeli doğruluğu 0,96'ya çıkmıştır. Bu modelin karışıklık matrisi ve tahminleme aşamasında kullandığı en önemli 15 tahminci ise Şekil 3.12'de verilmiştir. Bu sefer en önemli tahminci beklendiği gibi 'semptom no' çıkarken, 'kitle', 'şiş', 'ağrı', 'kus' ve 'ateş' gibi semptom isimleri en önemli tahminciler arasında yer almıştır. Bu model için her bir semptoma ait karışıklık matrisleri (Şekil 3.12) ve başarı ölçütleri hesaplanmıştır (Tablo 3.17). En başarılı tahmin sıfır hatayla 'kusma' için bulunmuştur. 'Ağrı' semptomu için 15 adet ağrı semptomundan 1 tanesi ağrı olarak belirlenememiştir. 174 adet kayıttan ise 2 tanesi ağrı semptomu olmadığı halde ağrı semptomu olarak tahmin edilmiştir. Ağrı semptomu olmadığı halde ağrı semptomu var olarak tahmin edilen 2 kaydın normalize edilmemiş (ham hali) aşağıdaki gibidir.

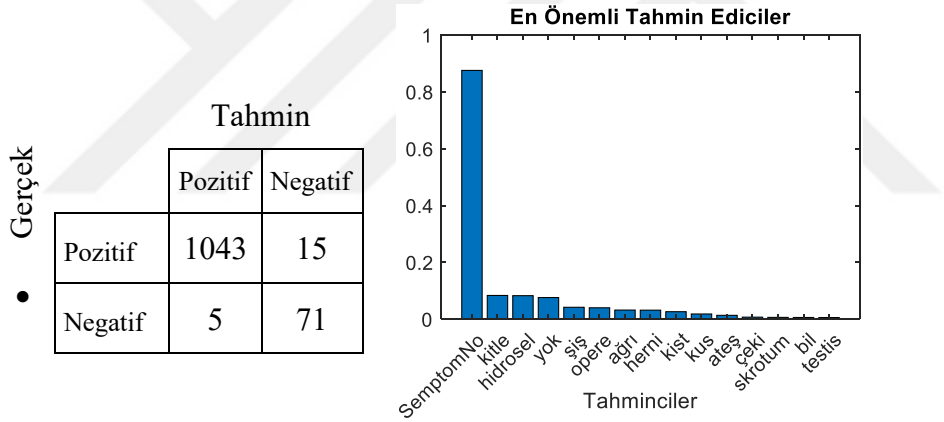
- "YAKLAŞIK 2 AY ÖNCE KASIK AĞRISI NEDENİ İLE YAPTIRDIKLARI TETKİKLER SONUCU ORTAYA ÇIKMIŞ. SAĞ

ALT POLDE KALKÜL ILMAY BİLGE VE EGEMEN GÖRMÜŞ. SAĞ GRADE 3-4 HN VAR. ŞU AN ŞİKAYET YOK. "

- "BOYUNDA ŞİŞLİK TAKİPLİ HASTA .HASTA 1 HAFTADIR ANTİBİYOTİK KULLANIYOR. LEZYONDA MUAYENE OLARAK MİNİMAL SINIRLANMA VE KÜÇÜLME VAR .ANCAK USG İLE BOYUTLARDA BÜYÜME SÖYLENİYOR. ESR GEÇEN HAFTAYA GÖRE 24'ten 45'E YÜKSELMİŞ.ŞUANDA PEDİATRİ TARAFINDAN İKİNCİ KÜR ANTİBİYOTERAPİ VERİLMİŞ. AGRI ,KIZARIKLIK ,ATEŞ GİBİ ŞİKAYETLERİ YOK."

Ağrı semptomu olmasına rağmen ağrı semptomunun belirlenemediği tek kayıt ise aşağıdaki gibidir.

- "1.5 AYDIR VARMIŞ BÜYÜMEMİŞ AĞRILIYMIŞ"



Şekil 3.13 MRMR Kullanıldığında 'Fine Tree' Sınıflandırıcı için Karışıklık Matrisi ve En Önemli 15 Tahminci

Tahmin

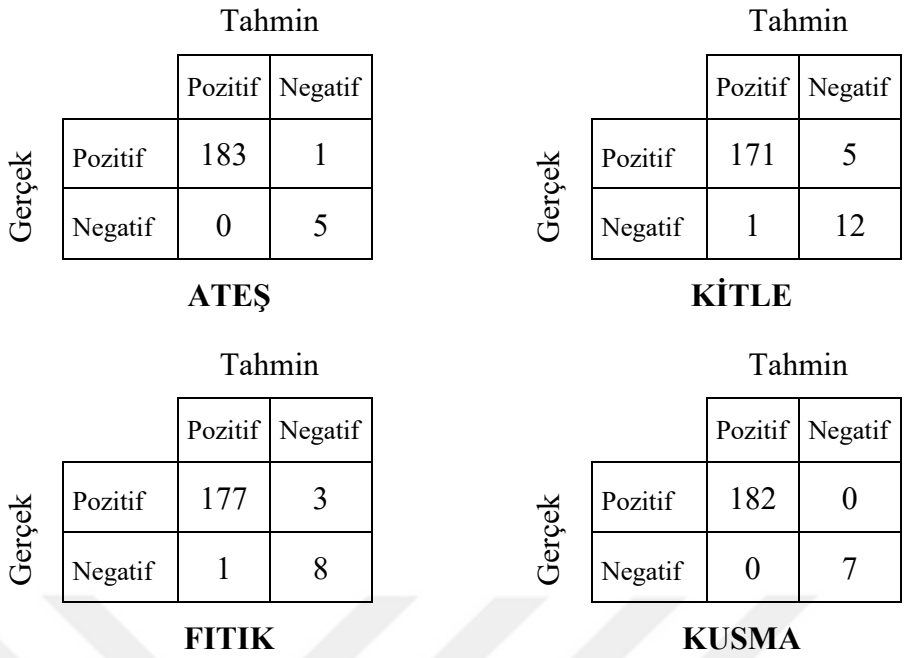
Gerçek	Pozitif	Negatif
Pozitif	172	2
Negatif	1	14

AĞRI

Tahmin

Gerçek	Pozitif	Negatif
Pozitif	158	4
Negatif	2	25

ŞİŞLİK



Şekil 3.14 MRMR Kullanıldığında 'Fine Tree' Sınıflandırıcı için Üretilen Model için Her Semptoma Karşı Düşen Karışıklık Matrisleri

Tablo 3.17 MRMR kullanıldığında 'Fine Tree' sınıflandırıcı için üretilen modelin test veri seti üzerindeki başarı ölçütleri

	Doğruluk	**Dengeli Doğruluk**	**Duyarlılık**	**Özgüllük**	**Kesinlik**	**F1 Skoru**
AĞRI	0,98	0,96	0,93	0,99	0,88	0,90
ŞİŞLİK	0,97	0,95	0,93	0,98	0,86	0,89
ATEŞ	0,99	1,00	1,00	0,99	0,83	0,91
KİTLE	0,97	0,95	0,92	0,97	0,71	0,80
FITIK	0,98	0,94	0,89	0,98	0,73	0,80
KUSMA	1,00	1,00	1,00	1,00	1,00	1,00

3.4.3.2 TF-IDF Yöntemi Kullanılarak Modellerin Eğitilmesi ve Test Edilmesi

TF-IDF yöntemi uygulandığında, kelime çantası yöntemine benzer şekilde, normalize edilmiş metinler 2146 elemandan oluşan sayısal vektörlere dönüştürülmüştür. Yine benzer şekilde bunlara 'semptom No' adlı değişken de eklendiğinde, veri seti 2147 adet öznitelik ve 1 çıktı değişkeninden oluşmaktadır.

Tablo 3.18 TF-IDF modeli için başarı ölçütleri

		Doğruluk	Dengeli Doğruluk	Duyarlılık	Özgüllük	Kesinlik	F1 Skoru	AUC
Fine Tree	Validasyon	0,98	0,92	0,84	0,99	0,90	0,87	0,91
	Test	0,98	0,94	0,89	0,98	0,79	0,84	0,95
Medium Tree	Validasyon	0,96	0,80	0,61	0,99	0,88	0,72	0,81
	Test	0,96	0,82	0,64	0,99	0,78	0,71	0,82
Coarse Tree	Validasyon	0,95	0,71	0,42	1,00	0,90	0,58	0,72
	Test	0,96	0,75	0,50	0,99	0,83	0,62	0,77
SVM (linear)	Validasyon	0,92	0,50	0,02	0,99	0,15	0,03	0,72
	Test	0,93	0,50	0,00	1,00	-	-	0,62
SVM (cubic)	Validasyon	0,91	0,51	0,03	0,99	0,20	0,05	0,79
	Test	0,93	0,50	0,00	1,00	-	-	0,86
SVM (coarse Gaussian)	Validasyon	0,92	0,50	0,00	1,00	-	-	0,75
	Test	0,93	0,50	0,00	1,00	-	-	0,66
Eff. Logistic Regression	Validasyon	0,92	0,52	0,04	0,99	0,38	0,07	0,78
	Test	0,93	0,52	0,04	1,00	0,75	0,08	0,76
Boosted Trees	Validasyon	0,97	0,84	0,68	1,00	0,94	0,79	0,98
	Test	0,98	0,88	0,78	0,99	0,84	0,81	0,99
Bagged Trees	Validasyon	0,98	0,92	0,85	0,99	0,90	0,88	0,97
	Test	0,97	0,92	0,86	0,98	0,77	0,81	0,99
RUSBoosted Trees	Validasyon	0,97	0,89	0,80	0,98	0,77	0,79	0,95
	Test	0,97	0,92	0,86	0,98	0,76	0,80	0,97
Wide Neural Network	Validasyon	0,95	0,78	0,59	0,98	0,73	0,65	0,82
	Test	0,94	0,80	0,64	0,96	0,53	0,58	0,84
Bilayered Neural Net.	Validasyon	0,93	0,72	0,47	0,97	0,57	0,51	0,78
	Test	0,92	0,73	0,50	0,95	0,43	0,46	0,79
Anahtar Sözcük Arama	Test	0,88	0,87	0,85	0,90	0,80	0,82	-

Kelime çantası yöntemi ile vektörleştirilen veri seti için yukarıda uygulanan makine öğrenmesi yaklaşımları TF-IDF yöntemi ile vektörleştirilen veri seti içinde de tekrarlanmıştır. Tablo 3.18'de farklı makine öğrenmesi için eğitilen modellere ait başarı ölçütleri verilmiştir. Tabloda görüldüğü gibi kelime çantası yöntemi ile benzer sonuçlar elde edilmiştir. Aşadaki tablolarda sırasıyla; maliyet matrisi kullanılarak, MRMR tekniği kullanılarak ve hem MRMR hem de maliyet matrisi kullanılarak elde edilen başarı ölçütleri en başarılı sınıflandırıcılar olan olan 'Fine Tree', 'Medium Tree', 'Boosted Trees', 'Bagged Trees' ve 'RUSBoosted Trees' için listelenmiştir. MRMR tekniğinde yine en önemli 1500 öznitelik seçilmiştir.

Tablolarda görüldüğü gibi kelime çantası yöntemi ile yine benzer sonuçlar elde edilmiştir.

Tablo 3.19 TF-IDF modeli için başarı ölçütleri (maliyet matrisli)

		Doğruluk	Dengeli Doğruluk	Duyarlılık	Özgüllük	Kesinlik	F1 Skoru	AUC
Fine Tree	Validasyon	0,97	0,93	0,88	0,98	0,78	0,83	0,91
	Test	0,97	0,93	0,88	0,97	0,71	0,78	0,88
Medium Tree	Validasyon	0,95	0,89	0,82	0,96	0,66	0,73	0,91
	Test	0,96	0,92	0,88	0,96	0,61	0,72	0,90
Boosted Trees	Validasyon	0,97	0,92	0,87	0,98	0,77	0,81	0,96
	Test	0,97	0,92	0,87	0,98	0,76	0,81	0,98
Bagged Trees	Validasyon	0,97	0,94	0,90	0,98	0,80	0,84	0,98
	Test	0,97	0,93	0,88	0,98	0,75	0,81	0,99
RUSBoosted Trees	Validasyon	0,97	0,90	0,82	0,98	0,76	0,79	0,95
	Test	0,97	0,94	0,91	0,98	0,74	0,82	0,97
Anahtar Sözcük Arama	Test	0,88	0,87	0,85	0,90	0,80	0,82	-

Tablo 3.20 TF-IDF modeli için MRMR kullanıldığındaki başarı ölçütleri

		Doğruluk	Dengeli Doğruluk	Duyarlılık	Özgüllük	Kesinlik	F1 Skoru	AUC
Fine Tree	Validasyon	0,98	0,91	0,83	0,99	0,88	0,86	0,93
	Test	0,98	0,96	0,93	0,99	0,83	0,88	0,93
Medium Tree	Validasyon	0,96	0,80	0,61	0,99	0,87	0,72	0,81
	Test	0,97	0,84	0,70	0,99	0,82	0,75	0,84
Boosted Trees	Validasyon	0,97	0,84	0,69	1,00	0,93	0,79	0,97
	Test	0,98	0,90	0,82	0,99	0,83	0,82	0,99
Bagged Trees	Validasyon	0,98	0,92	0,84	0,99	0,89	0,86	0,97
	Test	0,98	0,94	0,89	0,98	0,79	0,84	0,98
RUSBoosted Trees	Validasyon	0,97	0,89	0,80	0,98	0,78	0,79	0,95
	Test	0,97	0,94	0,91	0,98	0,74	0,82	0,97
Anahtar Sözcük Arama	Test	0,88	0,87	0,85	0,90	0,80	0,82	-

Tablo 3.21 TF-IDF modeli için MRMR kullanıldığındaki başarı ölçütleri (maliyet matrisli)

		Doğruluk	Dengeli Doğruluk	Duyarlılık	Özgüllük	Kesinlik	F1 Skoru	AUC
Fine Tree	Validasyon	0,97	0,93	0,89	0,98	0,77	0,82	0,92
	Test	0,97	0,95	0,93	0,97	0,70	0,80	0,93
Medium Tree	Validasyon	0,94	0,89	0,83	0,95	0,59	0,69	0,91
	Test	0,96	0,92	0,88	0,96	0,61	0,72	0,90
Boosted Trees	Validasyon	0,97	0,92	0,87	0,98	0,77	0,82	0,96
	Test	0,97	0,93	0,88	0,98	0,74	0,81	0,98
Bagged Trees	Validasyon	0,97	0,94	0,90	0,98	0,79	0,84	0,97
	Test	0,97	0,95	0,93	0,97	0,72	0,82	0,99
RUSBoosted Trees	Validasyon	0,97	0,89	0,81	0,98	0,77	0,79	0,95
	Test	0,97	0,94	0,91	0,98	0,74	0,82	0,96
Anahtar Sözcük Arama	Test	0,88	0,87	0,85	0,90	0,80	0,82	-

4
SONUÇ

Bu çalışmada, çocuk hastalara ait Türkçe klinik metinler içinden doğal dil işleme yaklaşımları kullanılarak otomatik semptom çıkarımı yapılmıştır. Bu semptomlar çeşitli hastalıkların tanısının belirlenmesi amacıyla makine öğrenmesi yöntemlerinde doğrudan öznitelik olarak kullanılabilir. Yapılan literatür araştırması ile bu alanda yapılan çalışmaların büyük bir kısmının İngilizce klinik metinler kullanılarak yapıldığı görülmüş ve Türkçe klinik metinler kullanılarak yapılan semptom çıkarımı çalışmasının ise bulunmadığı görülmüştür.

Çalışmada 2000 adet klinik metin, çocuk hastalıkları uzmanlarının yardımıyla etiketlenmiştir. Her ne kadar çok fazla sayıda semptom ayrıntılı bir şekilde etiketlenmiş olsa da semptom başına düşen kayıt sayısı düşük kaldığından, metinde en fazla etiketlenen ‘ağrı’, ‘şişlik’, ’ateş’, ’kitle’, ’fıtık’ ve ‘kusma’ semptomları sınıflandırma işleminde kullanılmıştır.

Veri setinde bulunan metin çok fazla sayıda teknik terim ve sadece belli kurumlarda çalışan uzmanların kullandığı kısaltma içermekteydi. Aynı zamanda bu metinler muayene esnasında çok kısa zamanda hızla yazıldığından dolayı çok fazla yazım yanlışı içermekteydi. Bu nedenle metnin ön-işlemesi ve özellikle normalizasyon işlemi oldukça zaman almıştır.

Normalize edilmiş metinleri sayısal özniteliklere dönüştürmek için kelime çantası, TF-IDF, N-gram ve FastText gibi yöntemler kullanılmıştır. Kelime çantası ve TF-IDF yöntemleri diğerlerine göre daha başarılı olduğu için bu yöntemlere ait analizler ayrıntılı bir şekilde verilmiştir. Vektörleştirme işleminin ardından veri seti eğitim ve test veri seti olmak üzere rastgele olacak şekilde ikiye ayrılmıştır. Makine öğrenmesi modellerinin eğitimi sırasında 10 kat çapraz doğrulama tekniği kullanılmıştır. Matlab Sınıflandırma Öğrenicisi (Classification Learner) araç kutusuyla çeşitli karar ağaçları, DVM, lojistik regresyon, yapay sinir ağları ve karar ağaçları için topluluk sınıflandırıcıları kullanılarak modeller eğitilmiş ve veri seti üzerinden test edilmiştir. Eğitilen modellerden çıkan sonuçlar karşılaştırıldığında en başarılı modeller bir karar ağacı olan ‘Fine Tree’ sınıflandırıcısı (doğruluk %98, dengeli doğruluk %94, AUC 0.95) ve topluluk sınıflandırıcıları olan ‘RUSBoosted

Trees' (doğruluk %97, dengeli doğruluk %92, AUC 0.97) ve 'Bagged Trees' (doğruluk %97, dengeli doğruluk %92, AUC 0.97) olmuştur. Kullanılan veri seti dengesiz olduğu için dengeli doğruluk ve AUC değerleri, doğruluk değerine göre daha güvenilirdir. Veri dengesizliğini ortadan kaldırmak için maliyet matrisindeki maliyet değerleri sınıf sayılarıyla ters orantılı olacak şekilde belirlenip, modeller yeniden eğitildiğinde, bir çok modelde dengeli doğruluk değeri kabaca %2-3 oranında artmıştır. Ayrıca bir öznitelik seçimi için geliştirilen MRMR yöntemi kullanıldığında 2147 adet öznitelikten oluşan veri seti 1500 adet özniteliğe indirilmiş ve makine öğrenmesi modellerinin eğitim maliyeti düşürülmüştür. 'Bagged Trees' sınıflandırıcısıyla eğitilen bu tip bir model (maliyet matrisiyle içeren) için %97 doğruluk, %96 dengeli doğruluk ve 0.98 AUC değerine ulaşılmıştır.

Bu çalışmada kullanılan az sayıda semptom için yüksek başarı oranları yakalanmıştır. Fakat çalışmada kullanılan semptom sayısının (semptomun şidddeti ve görüldüğü yerleri de içerecek şekilde) ciddi şekilde artırılmasına ihtiyaç vardır. Bunun için çok daha fazla verinin etiketlenmesi gerekmektedir. Burada elde edilen sonuçlar fazla sayıda kayıttan oluşan etiketli bir veri seti kullanılarak, çok daha fazla sayıda semptomun başarılıyla sınıflandırılabileceği konusunda umut vermektedir.

KAYNAKÇA

[1] Fua, Sunyang et al., “Clinical Concept Extraction: a Methodology Review,” Journal of Biomedical Informatics, vol. 109, no. 103526, 2020.

[2] Sager N, Friedman C, Lyman MS. Medical language processing: computer management of narrative data: Addison-Wesley Longman Publishing Co., Inc.; 1987.

[3] Y. Wang, L. Wang, M. Rastegar-Mojarad, S. Moon, F. Shen, N. Afzal, S. Liu, Y. Zeng, S. Mehrabi, S. Sohn, H. Liu, Clinical information extraction applications: a literature review, J. Biomed. Inform., 77, 34–49, 2018.

[4] X. Luo, P. Gandhi, S. Storey and K. Huang, "A Deep Language Model for Symptom Extraction From Clinical Text and its Application to Extract COVID-19 Symptoms From Social Media," in IEEE Journal of Biomedical and Health Informatics, vol. 26, no. 4, pp. 1737-1748, April 2022.

[5] A. R. Aronson, “Effective mapping of biomedical text to the UMLs metathesaurus: The metamap program,” in Proc. Amer. Med. Informat. Assoc. Symp., pp. 17–21, 2001.

[6] G. K. Savova et al., “Mayo clinical text analysis and knowledge extraction system (CTAKES): Architecture, component evaluation and applications,” J. Amer. Med. Informat. Assoc., vol. 17, no. 5, pp. 507–513, 2010.

[7] R. Vijayakrishnan et al., “Prevalence of heart failure signs and symptoms in a large primary care population identified through the use of text and data mining of the electronic health record,” J. Cardiac Failure, vol. 20, no. 7, pp. 459–464, 2014.

[8] R. G. Jackson et al., “Natural language processing to extract symptoms of severe mental illness from clinical text: The clinical record interactive search comprehensive data extraction (CRIS-CODE) project,” BMJ Open, vol. 6, no. e012012, pp. 1–10, 2017.

[9] A. V. Gundlapalli et al., “Detecting the presence of an indwelling urinary catheter and urinary symptoms in hospitalized patients using natural language processing,” J. Biomed. Informat., vol. 71, pp. S39–S45, 2017.

10] Van Putten, T.E., “Symptom extraction from atrial fibrillation patient clinical notes using deep learning,” Yüksek Lisans Tezi, Eindhoven University of Technology, 2020.

[11] J. M. Steinkamp, W. Bala, A. Sharma, and J. J. Kantrowitz, “Task definition, annotated dataset, and supervised natural language processing models for symptom extraction from unstructured clinical notes,” J. Biomed. Informat., vol. 102, Art. no. 103354, 2020.

[12] G. Divita, G. Luo, L. Tran, T. E. Workman, A. V. Gundlapalli, and M. H. Samore, “General symptom extraction from va electronic medical notes,” MEDINFO 2017: Precision Healthcare Through Informat., IOS Press, pp. 356–360, 2017.

[13] R. A. Stein, J. Patricia and J. F. Valiata, “An analysis of hierarchical text classification using word embeddings symptom” Information Sciences, cilt 471, pp. 216-232, 2019.

[14] Joulin, A., Grave, E., Bojanowski, P., & Mikolov, T. (2016). Bag of tricks for efficient text classification. arXiv preprint arXiv:1607.01759.

[15] R. Bellazzi and B. Zupan, “Predictive data mining in clinical medicine: Current issues and guidelines.” International Journal of Medical Informatics, vol. 77, no. 2, pp. 81-97, 2008, doi: 10.1016/j.ijmedinf.2006.11.006.

[16] https://code.google.com/archive/p/zemberek/, erişim tarihi: 26.06.2024.

[17] https://srm.metu.edu.tr/tr/system/files/tip-terimleri-sozlugu.pdf, erişim tarihi: 26.06.2024.

TEZDEN ÜRETİLMİŞ YAYINLAR

Konferans Bildirileri

1. Altunsaçan G., Aslanyürek B., Aydın E., Feature Extraction from Patient Data for Diagnosing Pediatric Patients, International Congress of New Searches in Sciences 2024, 8-9 June, İstanbul, Türkiye.

Projeler

1. Çocuk Hastaların Tanısının Belirlenmesinde Hasta Verilerindeki Özniteliklerin Çıkarılması, 2024, YTÜ BAP, Proje Kodu: TYL-2023-5673.