Journal of the Faculty of Engineering and Architecture of Gazi University 39:3 (2024) 1509-1523

Mühendislik Mimarlık Fakültesi Dergisi
Journal of The Faculty of Engineering and Architecture of Gazi University

Elektronik / Online ISSN : 1304 - 4915
Basılı / Printed ISSN : 1300 - 1884

Use of logistic regression in the diagnosis of thyroid cancer

Mehmet Emin Asan[1]*, Harun Taşkın[2], Murat Alemdar[3], Recayi Çapoğlu[4]

[1]Department of Industrial Engineering, Institute of Natural Sciences, Sakarya University, 54050, Serdivan, Sakarya, Türkiye
[2]Department of Industrial Engineering, Faculty of Engineering, Sakarya University, 54050, Serdivan, Sakarya, Türkiye
[3]Department of Neurology, Faculty of Medicine, Sakarya University, 54100, Sakarya, Türkiye
[4]Department of General Surgery, Faculty of Medicine, Sakarya University, 54100 Sakarya, Türkiye

Highlights:

Feature selection and data set creation using mathematical/statistical algorithms
• Use of machine learning algorithms in the diagnosis of thyroid cancer.
• • Use of logistic regression model in the diagnosis of thyroid cancer.

Keywords:

- Logistic regression
- Thyroid cancer
- Machine learning
- Medical diagnostics

Article Info:

Research Article
Received: 19.02.2023
Accepted: 11.08.2023

DOI:

10.17341/gazimmfd.1253193

Acknowledgement:

Correspondence:

Author: Mehmet Emin Asan
e-mail: masan@subu.edu.tr
phone: +90 553 177 6830

Graphical/Tabular Abstract

In this study, the most appropriate Machine Learning model was investigated in order to prevent unnecessary surgery for patients who were pre-diagnosed with thyroid cancer but whose pathology results were negative after surgery. With this research, the most appropriate classification model, modeled with the most appropriate data, was discussed. All the steps taken in the process of obtaining the model are shown in Figure A.

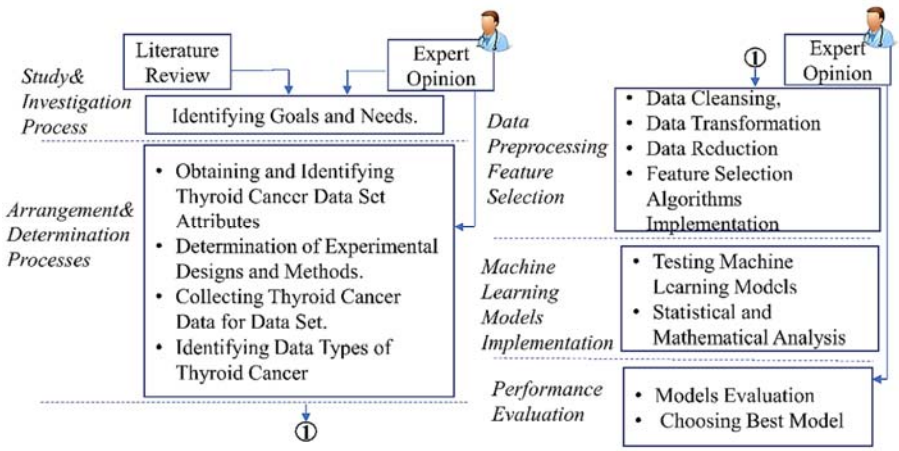


Figure A. Study Steps

Purpose: This study aims to obtain a near-accurate diagnosis for patients with thyroid cancer by using machine learning methods over previous patient data before the thyroid tissue is resected too much.

Theory and Methods: There is no clear answer as to what the best machine learning model will lead to for the data set at hand. However, converting the data diversity into appropriate data types, preparing data sets, scaling the data, selecting the features, and repeatedly testing with machine learning techniques/models can provide a satisfactory answer to the question asked above. The thyroid cancer data we have consists of continuous and discrete data types. The fact that the result data is binary and takes a binomial value necessitated the use of nonlinear methods. Machine learning models such as Naïve Bayes, Generalized Linear Model, Logistic Regression, Fast Large Margin, Deep Learning, Decision Tree and Random Forest were used for significance results.

Results: Cytology_Result_Benign and USG_LAP attributes pair, which we have tried among the 6 best features of twenty-three obtained using the logistic regression model, gives the best result of specificity with 99.31%.

Conclusion: According to the results of POST-OP (Final Operation) among those diagnosed with thyroid Cancer, 99.31 of every 100 people were correctly diagnosed. This result corresponds to 232.39 out of 234 cases. As a result, according to the data, the diagnosis of not having cancer was given incorrectly in only about 2 people based on the Logistic regression model we used. Accordingly, it was confirmed that the Cytology_result_benign and USG LAP attributes obtained in the outcome evaluation made with the specialists of the general surgery department of the Sakarya University research hospital are the attributes they consider in making a decision about performing the preoperative FNA test for the final.

Journal of the Faculty of Engineering and Architecture of Gazi University 39:3 (2024) 1509-1523

Mühendislik Mimarlık Fakültesi Dergisi
Journal of The Faculty of Engineering and Architecture of Gazi University

Elektronik / Online ISSN : 1304 - 4915
Basılı / Printed ISSN : 1300 - 1884

Tiroit kanseri hastalık tanısında lojistik regresyon kullanımı

Mehmet Emin Asan[1]*, Harun Taşkın[2], Murat Alemdar[3], Recayi Çapoğlu[4]
[1]Sakarya Üniversitesi, Fen Bilimleri Enstitüsü, Endüstri Mühendisliği, 54050, Serdivan, Sakarya, Türkiye
[2]Sakarya Üniversitesi, Mühendislik Fakültesi, Endüstri Mühendisliği Bölümü, 54050, Serdivan, Sakarya, Türkiye
[3] Sakarya Üniversitesi, Tıp Fakültesi, Nöroloji Anabilim Dalı, 54100, Sakarya, Türkiye
[4]Sakarya Üniversitesi, Tıp Fakültesi, Genel Cerrahi Bölümü, 54100 Sakarya, Türkiye

ÖNEÇIKANLAR

- Matematiksel ve/veya istatistiksel algoritmalar kullanarak öznitelik seçme ve veri seti oluşturma.
- Tiroit kanser hastalığı tanısında makine öğrenmesi algoritmalarının kullanımı
- Lojistik regresyon yönteminin tiroit kanser hastalığı tanısında kullanımı

Makale Bilgileri
Araştırma Makalesi
Geliş: 19.02.2023
Kabul: 11.08.2023

DOI:
10.17341/gazimmfd.1253193

Anahtar Kelimeler:
Lojistik regresyon, tiroit kanseri, k-en yakın komşu, destek vektör makineleri, rastgele orman, tıbbi tanı

ÖZ

Tiroit kanseri, 2020'deki sonuçlara göre, tüm kanserlerin küresel insidansının %3'üne karşılık gelirken bazı ülkelerde son 30 yılda önemli ölçüde artmıştır. Tiroit nodülü, tiroit bezinin içinde bulunan bir lezyondur. Bu lezyonların kanserli olma olasılığı önemli bir endişe kaynağıdır. USG ile saptanan nodüller 1 cm'den büyük ve kötü huylu olma (Malignant) açısından kuşkuluysa, ince iğne aspirasyon (İİA) biyopsisi kullanılır ve değerlendirmeler yapılır. İyi huylu İİA sonuçları, gereksiz tiroit ameliyatlarının önlenmesine yardımcı olur. Kötü huylu (Malign) hücreler tespit edilirse, İİA sonucu cerrahi stratejinin belirlenmesinde etkin bir faktör olur. Buna rağmen, cerrahlar kötü huylu hücre potansiyeline ilişkin belirsizlik nedeniyle, çok yüksek oranda iyi huylu (Benign) tiroit dokusu rezeke etmektedirler. Bu nedenle, daha doğru sonuçlar veren ve cerrahi işlem gerektirmeyen (non-invasive) tekniklere ihtiyaç duyulmaktadır. Bu çalışmanın amacı, tiroit dokusu çok fazla rezeke edilmeden, hastanın verileri üzerinden makine öğrenmesi metotlarından biri olan Lojistik regresyon kullanarak, kesine yakın tanının elde edilmesidir. Bu çalışma ile hastaların test sonuçlarını kullanarak, nodülün kötü huylu (kanserli) olup olmadığını tahmin eden bir model üzerinde denemeler yaptık. Gerçekte kanserli olmadığı halde operasyon geçiren hasta sayısı üzerinde spesifik olarak yoğunlaşarak özgüllük (specificity) analizi yaptık. Lojistik regresyon sınıflandırma algoritması ile elde edilen sonuçlar içerisinden en iyi spesifiklik/özgüllük değerini %99,31 olarak elde ettik.

Use of logistic regression in the diagnosis of thyroid cancer.

HIGHLIGHTS

- Feature selection and data set creation using mathematical and/or statistical algorithms
- Use of machine learning algorithms in the diagnosis of thyroid cancer disease
- Use of logistic regression method in the diagnosis of thyroid cancer

Article Info
Research Article
Received: 19.02.2023
Accepted: 11.08.2023

DOI:
10.17341/gazimmfd.1253193

Keywords:
Logistic regression, thyroid cancer, k-nearest neighbor, support vector machines, random forest, medical diagnosis

ABSTRACT

While thyroid cancer accounts for 3% of the global incidence of all cancers, according to results in 2020, it has increased significantly in some countries over the last 30 years. The possibility that the thyroid nodule is cancerous is a significant concern. If nodules detected by USG are larger than 1 cm and are suspicious for malignancy, they are evaluated with fine needle aspiration (FNA) biopsy. Benign FNA results help prevent thyroid surgeries. If malignant cells are detected, it becomes an effective factor in the surgical decision. Surgeons resect a very high percentage of benign thyroid tissue due to uncertainty regarding the potential for malignant cells. Therefore, techniques that provide more accurate results and do not require surgical procedures are needed. The aim of this study is to obtain a near-definitive diagnosis by using Logistic regression, one of the machine learning methods, on the patient's data, without resecting the thyroid tissue too much. In this study, we experimented with a model that predicts whether the nodule is malignant (cancerous) or not, using the test results of the patients. We conducted a specificity analysis by focusing specifically on the number of patients who underwent surgery even though they did not have cancer. Among the results obtained with the logistic regression classification algorithm, we obtained the best specificity value of 99.31%.

*Sorumlu Yazar/Yazarlar / Corresponding Author/Authors : *masan@subu.edu.tr, taskin@sakarya.edu.tr, muratalemdar@sakarya.edu.tr, rcapoglu@gmail.com / Tel: +90 553 177 6830

1. Giriş (Introduction)

Farklı ülkelerde farklı yorumlansa da *Tıp Bilimi* ile *Tıp Uygulamaları* ayrı alanlar olarak ele alınmakta ve Tıp uygulamaları kliniklerde daha geniş ve daha yaygın bir biçimde gerçekleştirilmektedir. Uygulamaların sürekli olması, klinik çalışmalarının ön planda olmasının en önemli nedenlerinden biridir.

Hizmetin kısa sürede verilmesi, hizmetin doğruluğunu teyit edememektedir. Özellikle hastaların hastalık belirtilerini net olarak aktaramamaları, tanı sonuçlarını olumsuz etkilerken, tanı için verilerin tahlil ve tetkikler üzerinden elde edilmesi ve yorumlanması ise uzman tecrübesi gerektirmektedir.

Hastalık tanısı için uzmanlar tarafından başvurulan uygulamalar arasında endoskopi, ultrason, sintigrafi, tomografi, manyetik rezonans, röntgen yer alırken kan testleri de önemli yer tutmaktadır. Ön tanı üzerinden gelişen süreç, daha hassas ve kesin tanının yapılabilmesi ve ihtisaslaşan tıbbi alanlar içinde en çok hangisine dahil olabileceği ile ilgili kararlar için ileri muayene ve tetkiklere ihtiyaç duyar.

Birçok hastalık gibi kanserli hastalar da kesin tanı konulana ve tedavi yöntemi belirlenene kadar geçen sürede çeşitli testlere ve tetkiklere tabi tutulurlar. Ön tanı yoluyla kanser şüphesi oluşan hastalar için çeşitli tetkik sonuçları değerlendirilmekte ve verilen son tanı kararı sonrası tedavi süreci başlatılmaktadır.

Kanser hastalığı genellikle vücutta hızlı yayılan kötü huylu tümörlerin varlığı ile ifade edilir. Tümörlerin çokluğu ve yayılma hızı kanserin yıkıcılığı ve ciddiyeti hakkında bilgi verir. Teşhisinde geç kalınmış ileri evre kanser hastalığı öldürücü olmaktadır [1].

Kanser hastalıklarından biri olan tiroit kanseri boğaz bölgesindeki tiroit bezi dokularından gelişen bir kanser çeşididir. Bu bölgede hücreler aşırı hızlı büyüyerek vücudun diğer bölgelerine yayılma potansiyeli gösterirler. Boğaz kısmında şişlik, nefes almada ve yutkunmada zorluk ve sesin anormalleşmesi gibi birçok belirtisi vardır. Tanı öncesi boğazın elle kontrolü ve aile hastalık geçmişinin analizi yapılır. Tiroit testleri, kan testleri ve görüntülü tetkikler incelenir. İnce iğne aspirasyon (İİA) adı verilen yöntem ile alınan hücrelerin biyopsi sonuçları değerlendirilir [2].

Tiroit kanseri, 2020'de tahmin edilen 586.000 yeni hasta ile tüm kanserlerin küresel insidansının %3'ünü temsil etmektedir [3]. Artış oranları, popülasyonlar arasında değişiklik gösterse de birkaç yüksek ve orta gelirli ülkelerde tiroit kanseri insidansı son 30 yılda önemli ölçüde artmıştır [4-5]. Artan sayıda tiroit kanseri teşhisi, büyük ölçüde ve genel olarak sağlık hizmetlerine erişimin artmasıyla birlikte tanısal görüntüleme teknolojisinin ve tıbbi gözetimin giderek yaygınlaşması ile ilişkilendirilebilir [6]. Artan insidansın aksine, tiroit kanseri için ölüm oranları azalmış veya sabit kalmıştır ve hemen hemen her yerde düşük seviyelerdedir [7].

Tiroit nodülü, tiroit bezinin içinde kendisini çevreleyen tiroit parankiminden radyolojik olarak ayırt edilebilen bir lezyondur. Nodüller dokunmakla anlaşılabilir ya da tesadüfi olarak saptanabilir. Erişkinlerin yaklaşık %60'ında bir veya daha fazla tiroit nodülü bulunur. Kanser olasılığı en önemli endişe kaynağıdır, ancak yalnızca yaklaşık %5inin kötü huylu olduğu kanıtlanmıştır [8]. Tiroit nodüllerine yaklaşımda fizik muayene, anamnez, serum tiroit fonksiyon testleri, ultrasonografi (USG) kullanılır. USG ile saptanan nodüller 1 cm'den büyük ve kötü huylu olma açısından kuşkuluysa, ince iğne aspirasyon (İİA) biyopsisi alınır ve değerlendirilir. İİA biyopsisi, tiroit nodülleri için altın standart tanı aracı olarak kabul edilir. İyi huylu İİA sonuçları, gereksiz tiroit ameliyatlarının önlenmesine yardımcı olur. Kanserli hücreler tespit edilirse, İİA sonucu, cerrahi stratejinin belirlenmesinde belirleyici bir faktördür [9]. Buna rağmen cerrahlar, kanserli hücre olma potansiyeline ilişkin belirsizlik nedeniyle çok yüksek oranda iyi huylu (kansersiz) tiroit dokusu rezeke etmektedir [10]. Bu nedenle daha doğru sonuçlar veren ve cerrahi işlem gerektirmeyen (non-invasive) tekniklere ihtiyaç duyulmaktadır.

Sonuçlar, elde edilen verilerin tam olarak doğru olmaları kaydıyla, sistematik bir algoritmaya tabi tutulmaması halinde farklı doktorlar tarafından farklı yorumlanabilmekte ve farklı tanılar konulabilmektedir. Bu durum, tıp alanında genelleme yapıldığında, daha iyi çözümlerin elde edilmesi sürecinde kaybedilmesi muhtemel zaman ve para maliyeti anlamına gelmektedir.

Birçok alanda kullanıldığı gibi tıp alanında da tanı koymak üzere karar destek sistemleri için yöneylem araştırması, istatistiki hesaplamalar ve matematiksel modeller üzerinden çözümler kullanılmaktadır. Tüm bu yöntemlerle birlikte "bilişsel yöntemler" kullanılsa da karar için uygun parametreler elde edilmeden ve bu parametreleri yorumlayacak yeterli tecrübe oluşturulmadan elde edilecek tanı sonuçları yanıltıcı olmaktadır.

Herhangi bir tanı için uygun verilerin bir araya getirilmesi ve veriler arasında en iyi ve en uygun bağlantının kurulması, hedeflenen ilk yöntem olmalıdır. Kesine yakın tanının konulması için en önemli iki parametreden biri verilerin tutarlılığı iken, bir diğeri tutarlı olan verileri birbiriyle ilişkilendirecek ve buradan bir sonuç çıkaracak uzmanın tecrübesi olmaktadır.

Uzman doktor, pratisyen hekim, hemşire, sağlık memuru ve diğer tüm sağlık personeli göz önüne alındığında, hedefe varmak konusunda insani acelecilik, bilişsel yöntemlerin mesleki uzmanlık tam olarak oluşmadan her seviyede kullanılması, insan oluşumuzdan kaynaklanan yorgunluk ve bıkkınlık, yanlış meslek seçimi neticesinde işin sevilememesi, tetkik/tahlil incelemelerinde yeterli zamanın olmayışı türünden birçok sebeple uygun tanı ve tedavi yapılamamaktadır. Bu duruma bazen, uzman başına düşen hasta sayısının fazlalığı ve tetkik için yeterli materyalin olmayışı da eşlik edebilmektedir.

Tıbbi tanı koymak için kliniklerde istenilen tahlil ve tetkikler çok sayıda veri içerse de veri içerisinden gerekli bilginin tam olarak çekilip çıkarılamaması, istenilen sonuçlar açısından verim ve kaliteyi olumsuz etkilemektedir. Bu veriler, çoğunlukla bir uzman doktor tarafından kişisel olarak doğrusal ve analitik yöntemlerle çözülmeye çalışılmaktadır.

İhtisaslaşma ile birlikte verilerin çoğalması, eldeki verilerin yorumlanması adına daha fazla zamana ihtiyaç duyulmasına neden olmaktadır. Eldeki veriler arasında bağlantının varlığı ve derecesi hakkında daha fazla fikir öne sürmek gerekliliği ortaya çıkmaktadır. Bu gereklilikler, tıp uygulamalarının verimliliği adına tıp biliminin önünü hızla açmaktadır.

Eldeki verilerin doğru tanı ve tedavi için anlamlı birer bilgiye dönüştürülmesi çabası, beraberinde yeni tekniklere ve modellere ihtiyacı doğurmaktadır. Genelde mühendislik, özelde ise Yapay zekâ teknikleri olarak ortaya çıkan bu yöntemlerin, doktorların tanı koymalarına yardımcı olması ve maliyetleri azaltması öngörülmektedir.

Veri bilimleri olarak ön plana çıkan bilim dalı, elde edilen verilerden anlamlı bilgiler çıkarmak ve elde edilmek istenen sonuca destek

vermek amacıyla Veri madenciliği ve Makine öğrenmesi tekniklerini birlikte kullanır. Uygun formatta ve uygun veri tipleri ile oluşturulan veri setleri, çeşitli makine öğrenmesi algoritmaları yoluyla birbiri ile ilişkilendirilerek gerekli bilgi elde edilebilmektedir.

Karar destek sistemi olarak kullanılmakta olan teknolojik gelişmeler, son zamanlarda Veri madenciliği ve Makine öğrenmesi üzerinden hız kazanmaktadır. Birçok farklı sektörde uygulanmasının yanı sıra, tıpta geleneksel tanı ve tedavi yöntemlerinin verdiği sonuçlardan daha iyi sonuçlar veren araştırmalar yapılmış ve sonuçlar elde edilmiştir [11]. Eldeki hasta verileri üzerinden, Veri madenciliği ve Makine öğrenmesi yöntemleri kullanılarak en az hata oranı ile tanı yapılabildiği ortaya konmuştur [12]. Veri madenciliği yöntemleri ve Makine öğrenmesi teknikleri birlikte çok daha karmaşık problemler çözümleyebilmiş ve veriler arası ilişkiler doğrusal veya doğrusal olmayan algoritmalarla açığa çıkarılmıştır [13].

Yapay zekâya ve alt alanlarına dayalı sistemlerin oluşturulması ve mevcut geleneksel sağlık sistemine entegre edilmesi, birtakım tereddütleri beraberinde getirse de maliyetin ve tanıya götüren sürecin kısaltılması, tanının daha yüksek oranda ve kısa sürede doğru yapılması, ölüm oranlarının azalmasına etki etmesi ve benzeri birçok nedenle gerekli görülmektedir [14].

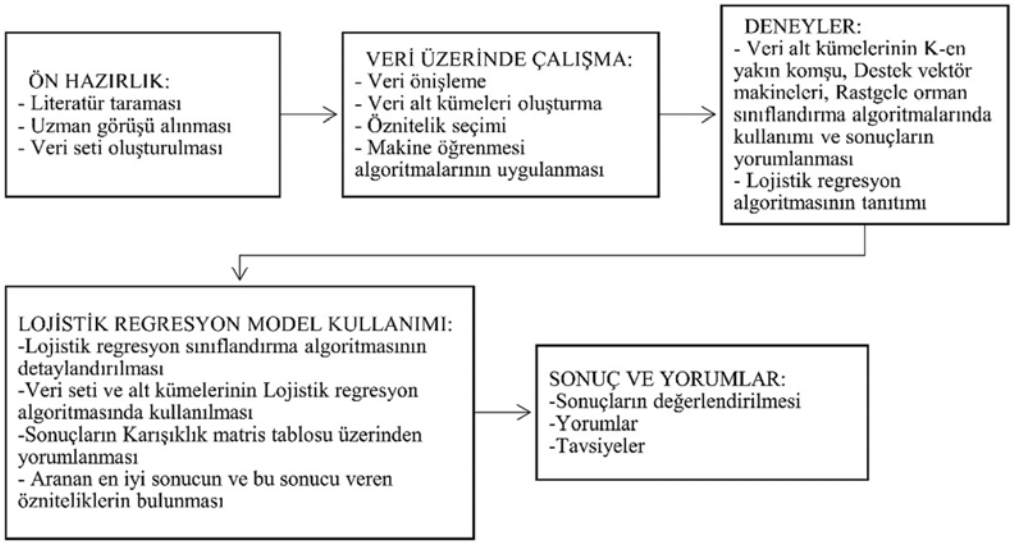
Veri madenciliği yaklaşımı ile birlikte Yapay zekâ, sağlık alanlarında tanı ve tedavide en iyi sonuçlara varmak üzere verileri edinme, işleme ve kullanıcıyı en iyi şekilde bilgilendirme yetilerine sahiptir. Yapay Zekâ, belirtilen işlemleri Makine öğrenmesi ve Derin öğrenme algoritmaları üzerinden gerçekleştirmektedir. Genelde Yapay zekâ özelde Makine öğrenmesi ve Derin öğrenme algoritmaları insan zekasını taklit ederler. Makine öğrenmesi, girdi verileri ile çıktı verileri arasında sebep sonuç ilişkileri kurar ve veriler arası ilişkileri göz önüne alarak verilen sonuçlar üzerinden kalıplar oluşturarak çeşitli algoritmalar yolu ile öğrenmesini tamamlar.

Algoritmalar, veriler arası ilişkiler üzerinden kesine yakın sonuçlar elde etmek üzere tahminde bulunabilirler ancak sonuçların nedenleri üzerinde tahmin yürütemezler. Gerçek hayat senaryoları ile çatışması muhtemel matematiksel ve istatistiksel analizlerin içine düşebileceği muhtemel girdaplar da hesaba katılmalı ve algoritmalar iyi hazırlanmış veri setleri üzerinde tekraren çalıştırılmalıdırlar [15]. Bu makalede tiroit kanseri ön tanısı konulan hastaların kan testlerini, USG, İİA biyopsisi sonuçlarını veri olarak kullanarak, nodülün kötü huylu (kanserli) olup olmadığını tahmin eden bir model üzerinde çalıştık. Üzerinde çalışılacak Makine öğrenmesi modeline karar vermeden önce, eldeki veriler farklı modellerde kullanılmak amacıyla hazır hale getirildi. Bu hazırlık, ikinci bölümde "Veri setinin oluşturulması ve hazırlanması" başlığı altında işlendi. Üçüncü bölümde, Makine öğrenmesi algoritmaları üzerinden son zamanlarda yapılmış tanı tahmini çalışmaları yanı sıra, Lojistik regresyon modeli tanımlaması üzerinde duruldu. Dördüncü bölümde, elimizde var olan tiroit kanseri hasta veri seti ve çeşitli öznitelik seçme testleri ile elde edilen veri seti alt kümeleri, K-en yakın komşu, Destek vektör makineleri ve Rastgele orman gibi sınıflandırma algoritmalarında denendi ve spesifiklik (specificity) üzerinden en iyi sonucu veren Lojistik regresyon modeli karışıklık matris tablosu ile yorumlandı. Beşinci bölümde ise sonuçlar değerlendirilerek yorumlamalar yapıldı. Makalede yapılan işlemlerin akışını ve atılan adımları içeren şema Şekil 1'de verilmiştir.

2. Veri Setinin Oluşturulması ve Hazırlanması
(Creation and Preparation of the Data Set)

Makine Öğrenimi büyük ölçüde verilere bağlıdır. Algoritmaların eğitilmesini mümkün kılan, makine öğreniminin ilerlemesinde ve çokça tercih edilmesindeki en önemli husus budur. Ancak, gerçekte veri biliminde uzmanlık seviyeniz ne olursa olsun verilerden anlam çıkaramıyorsanız, kapasitesi ne olursa olsun hiç bir yazılım ve/veya makine işe yaramaz.

Alelade ve karışık bir halde elde edilen tüm veri kümeleri kusurlar içerir. Bu veri kümelerinin çoğu işlenmemiş, tutarsız, eksik ve gürültülü veriler içerdiklerinden analiz edilmeleri zordur. Bu veri setleri, kullanılmadan önce kaliteli ve işe yarar hale getirilmelidirler. Veri ön işleme burada, verilerin tutarlılığını ve kalitesini sağlayan bilgi keşif sürecinin bir adımı olarak karşımıza çıkar [16]. Bu makalede kullanılan veri seti, Sakarya Üniversitesi Araştırma Hastanesi Genel Cerrahi Bölümünden alınan 234 adet tiroit kanserli hasta verisinden oluşmaktadır. Sakarya Üniversitesi Etik Kurul kararı üzerinden gerekli izinler alındı. Verilerin veri tiplerinin belirlenmesi



Şekil 1. İş akış adımları ve yapılan işlemler (Workflow steps and operations performed.)

için öncelikle tiroit kanseri operasyonlarını gerçekleştiren uzmanlarla bir araya gelindi. Modellemeler için gerekli en iyi öznitelik seçimi adımları dışında kalan tüm diğer veri ön hazırlık adımları, verilerin sistemden çekilmesi sırasında uzman kontrolünde elle tek tek gerçekleştirildi. Tiroit kanserli hasta verileri içindeki tüm öznitelikler, anlamları ve veri tipleri ile birlikte Tablo 1'de bir arada verilmiştir. Tiroit kanseri hasta veri seti içinden en iyi tanı sonucunu verecek öznitelik kümesini elde etmek üzere çeşitli yöntem ve algoritmalar kullanılmadan önce; tiroit kanseri veri seti, farklı Normalleştirme (Normalization) ve Ölçeklendirme (Scaling) yöntemlerine tabi tutularak farklı veri setleri oluşturuldu. Bu adımda Mutlak maksimum ölçeklendirme (Absolute maximum scaling), Min-maks ölçeklendirme (Min-max scaling), Ortalama (mean) kullanarak Normalleştirme (Normalization) , Güçlü ölçeklendirme (Robust Scaling) ve Standartlaştırma (Standardization) yöntemleri kullanılmıştır. Bu işlemlerin amacı, farklı ölçeklendirme yöntemlerinin öznitelik seçimi ve sonuçlara etkisini öğrenmektir. Bu aşamada, henüz alt veri grupları oluşturulmamıştır. Veri seti oluşturmak üzere takip edilen adımlar, Tablo 2'de gösterildiği gibi gerçekleştirilmiştir.

Tablo 1. Kullanılan tiroit kanseri veri setine ait verilerin anlamı ve veri tipleri
(The meaning and types of the data belonging to the thyroid cancer data set used.)

Öznitelik	Anlamı	Orijinal veri tipi ve içerik	Veri tipi (Dönüştürme sonrası)
Cinsiyet	Kadın ya da Erkek	İki terimli (Binominal:Kadın:1 Erkek: 2)	Sayısal-ikili (0,1)
Yaş	Yaş	Sürekli (20-87 yaş aralığı)	Sayısal-Sürekli
Usg	Ultrasonografi	Çok terimli (Polinominal, kötü huylu, olası kötü huylu, iyi huylu, olası iyi huylu)	Sayısal-kategorik (0,1,2,3)
Usg Lap	Ultrasonografi ile Lenfadenopati	İki terimli (Binominal , yok, mevcut)	Sayısal-ikili (0,1)
Şüpheli Nodül Çapı	Kanser şüpheli nodülün çapı	Çok terimli (Polinominal ,1 cmden küçük,1-2 cm arası,2-3 cm arası,3-4 cm arası,4 cmden büyük)	Sayısal-kategorik (0,1,2,3,4)
Boy/En Oranı	Kanserden şüphe edilen nodülün boy/en oranı	İki terimli (Binominal, Genişliğinden daha uzun olan, Uzunluğu genişliğinden az)	Sayısal-ikili (0,1)
Nodül Ekosu	Nodülün izoekoik, hipoekoik olma durumu	Çok terimli (Polinominal, İzoekoik, Hipoekoik, Hiperekoik)	Sayısal-kategorik (0,1,2)
Mikro kalsifikasyon	Göğüs dokusunda rastgele oluşan kalsiyum(kireç) çökeltisi.	İki terimli (Binominal, mevcut, yok)	Sayısal-ikili (0,1)
Nodül Özelliği	Solid ya da kistik	İki terimli (Binominal, solid, kistik)	Sayısal-ikili (0,1)
Sitoloji sonucu	Operasyon öncesi sonuç	Çok terimli (Polinominal, kötü huylu, olası kötü huylu, iyi huylu)	Sayısal-kategorik (0,1,2)
Tiroit fonksiyonu	Fonksiyonel sonuç	Çok terimli (Polinominal, Hipotroidi, hipertroidi, normal)	Sayısal-kategorik (0,1,2)
Tsh	Tiroit uyarıcı hormon kan testi	0-76,7 arası ondalık sayılar	Sayısal-Ondalık
sT3	Kandaki sT3	2,53-13,6 arası ondalık sayılar	Sayısal-Ondalık
sT4	Kandaki sT4	2,38-41,05 arası ondalık sayılar	Sayısal-Ondalık
Beyaz Küre	Lökosit.	3,43-14,0 arası ondalık sayılar	Sayısal-Ondalık
Rbc	Alyuvar hücreleri	3,67-7,16 arası ondalık sayılar	Sayısal-Ondalık
Hgb	Hemoglobin	8,3-16,9 arası ondalık sayılar	Sayısal-Ondalık
Platelet	Pıhtılaşmayı sağlayan hücre	83,10-2500 arası ondalık sayılar	Sayısal-Ondalık
Mcv	Kırmızı kan hücre boyutu.	29-99,10 arası ondalık sayılar	Sayısal-Ondalık
MCHC	Ortalama hemoglobin	27,24-37,8 arası ondalık sayılar	Sayısal-Ondalık
Mentzer_indeksi	Mcv/rbc oranıdır.	6,04-25,72 arası ondalık sayılar	Sayısal-Ondalık
Neutrophil	Beyaz kan hücresi.	1,62-11,9 arası ondalık sayılar	Sayısal-Ondalık
Lenfosit	Beyaz kan hücresi.	0,125-4,25 arası ondalık sayılar	Sayısal-Ondalık
postopSonuc	Operasyon sonrası kesin sonuç.	İki terimli (Binominal, Kansersiz, Kanserli)	Sayısal-ikili (0,1)

Tablo 2. Hasta verileri toplama adımları ve amaçları (Patient data collection steps and their purposes.)

Adımlar	Yapılan işlemin amacı
Ön teşhis öncesi istenilen testlerin ve hasta bilgilerinin belirlenmesi	Operasyon öncesi teşhis koymak amacıyla test sonuçlarının veri seti içinde kullanılması.
Öznitelik adlarının oluşturulması	Teşhis amacıyla gerekli test sonuçlarının veri setinde değişken olarak kullanılmak üzere adlandırılmaları (Genellikle yapılan işlemle aynı adı taşır. Kısaltmalar kullanılabilir.)
Öznitelik veri tipi belirleme	Uygun algoritmalara tabi tutmak amacıyla özniteliklerin değişken özelliklerini açığa çıkarmak.
Veri tiplerinin Sayısallaştırılması	Kullanmakta olduğumuz veri setinin farklı matematiksel ve/veya istatistiksel algoritmalarda kullanımını kolaylaştırmak.

8Tiroit kanseri veri seti oluştururken kullanılan tahlil/test sonuçlarının tamamı, sayısal olarak mevcut değildir. Sonuçların bazıları evet/hayır şeklinde iken, örneğin bir başka test olan USG sonucu kötü huylu, olası kötü huylu, iyi huylu ve olası iyi huylu olmak üzere dört adet çok terimli (polynomial) değerler alabilmektedir. Öte yandan, nominal değerlerin yanında farklı ondalık sayılardan oluşan ve veri tipi *sürekli* olan öznitelikler de veri setimiz içinde yer almaktadır. Veri setimiz içinde ham olarak gelen verilerin öznitelik adlandırmaları, anlamları, veri tipi ilk görünümleri ve farklı tipe dönüştürülmüş halleri Tablo 1'de gösterilmiştir.

2.1. Veri Önişleme (Data Preprocessing)

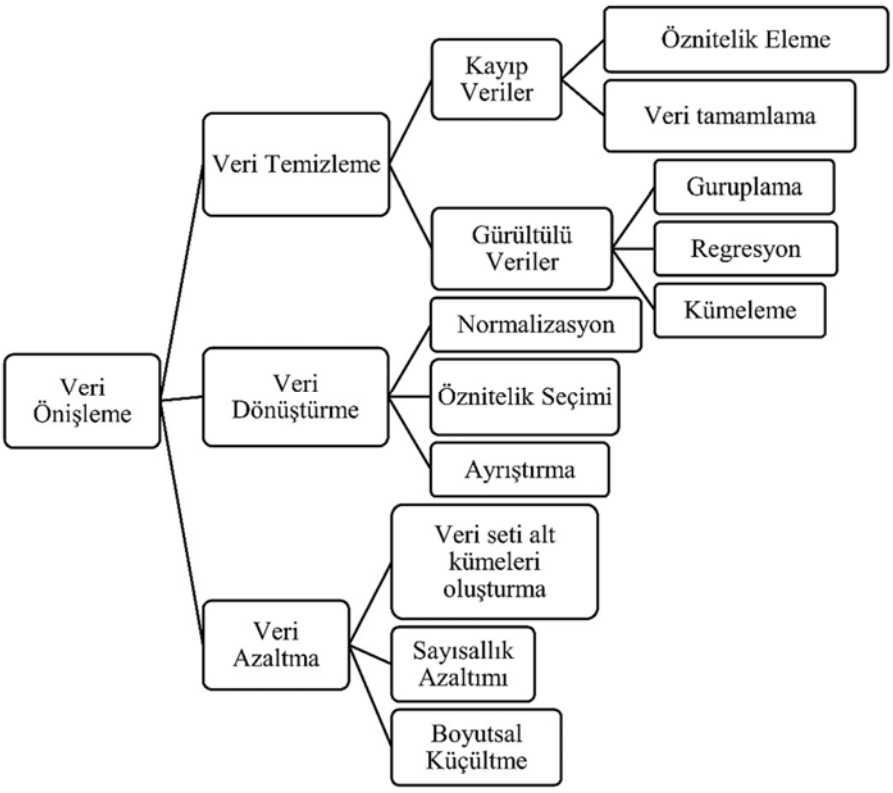
İstenen en iyi sonuçları elde etmek için, veri tiplerine uygun Algoritmaların-modellerin bulunması çok önemli olmakla birlikte, verilerin veri setleri halinde çok iyi bir ön hazırlığa tabi tutulmaları gerekmektedir. Veri biliminde "Veri önişleme" ya da "Veri ön hazırlığı" adıyla yer alan ve verileri hazır hale getirmek için gerekli işlemler, hedeflenen sonuçları doğrudan etkilemektedir. Kurallara uygun ve doğru şekilde hazırlanan veriler, kullanılmak istenilen Makine öğrenmesi modellerinin doğru sonuçlar vermesini sağlayacaktır. Veri ön işleme, hangi algoritma ya da model üzerinde çalışıldığına bakılmaksızın en iyi sonuçları elde etmek amacına yönelik önem arz etmektedir ve üzerinde dikkatlice çalışılmalıdır. Eldeki veri seti Makine öğrenmesi modellerine girdi olarak verilmeden hemen önce, kullandığımız verileri veri setleri halinde hazır hale getirmek amacıyla ön işleme tabi tutmak üzere Tablo 3'te verilen temel işlemler takip edilmiştir.

Çeşitli Veri önişleme teknikleri mevcuttur. Veri önişleme işlemlerinde her bir adım farklı yöntem ve/veya algoritmalarla yapılabilmektedir. Veri önişleme temel adımlarımız Şekil 2'de gösterilmiştir.

En uygun modelin bulunması amacıyla hazırlanan veri setleri, Şekil 2'de gösterilen temel adımlardan geçirilmelidir. Bu adımların verilere uygulanması için, çeşitli yöntem ve teknikler mevcuttur. Ön hazırlık yapılırken, ilk adım verilerin temizlenmesi, tamamlanması ve eksiği

Tablo 3. Veriler için ön hazırlık yapma amaçları ve gerçekleştirilen işlemler.
(Purposes of preprocessing of data and activities carried out.)

Amaç	Gerçekleştirilen İşlemler.
Veride kesinliğin (preciseness of information) elde edilmesi	Doğru/Yanlış verilerin birbirinden ayrılması/ayıklanması
Verilerin tam olması (Completeness)	Kaydedilmemiş ya da ulaşılamayan verilerin bulunup kaydedilmesi
Tutarlılığın (Consistency) sağlanması	Veri setinin kurallara uygun olarak oluşturulması
Veri Güncelleme	Güncel olmayan verilerin güncellenmesi
Verinin kolay anlaşılabilmesi	Anlaşılır ve yorumlanabilir düzeye getirilmesi için; • Verinin temizlenmesi • Verinin entegrasyonu • Verinin küçültülmesi • Verinin dönüştürülmesi • Verinin ayrıklaştırılması



Şekil 2. Veri önişleme adımları (Data preprocessing steps)

fazla olan özniteliğin setten çıkarılması işlemlerinden oluşmaktadır. İlk adım gerçekleştirildikten sonra, veriler arası ilişki düzeylerinin ayarlanması ve aynı seviyede işlem görebilmeleri için veri dönüştürme işlemleri gerekmektedir. Bu aşamada, tüm veri seti içinden, sonucu verecek en iyi özniteliklerin seçimi yapılır. Son aşamada, eğer gerekli ise veri setleri en uygun alt kümelere ayrıştırılarak en iyi veri alt kümeleri elde edilir.

2.1.1. Veri temizleme (Data cleansing)

Genelde tüm Veri madenciliği modellerinde, özelde ise aynı yol ve yöntemleri kullanan Makine öğrenmesi modellerinde kesine yakın sonuçların alınabilmesi için kullanılan tüm verilerin, veri tabanı yönetim kurallarına göre temizlenerek hazır hale getirilmeleri gerekmektedir. Temizlemek, gürültülü verilerin giderilmesinin yanı sıra, kayıp verilerin ortalama (mean) ya da orta eleman alma (median) türünden istatistiki yaklaşımlarla tamamlanmalarını da içermektedir. Çeşitli gruplama, sınıflama veya kümeleme teknikleri ile test edilerek grafiklerle rahatça gösterilebilen aykırı ve gürültülü veriler, veri setinden temizlenir ya da kaynak ile karşılaştırılarak revize edilirler. Eğer herhangi bir öznitelik sütunu içinde, hastalara ait veriler çok fazla eksikse ya da tümü aynı veriye sahipse, ilgili özniteliğin tamamen setten çıkarılması yararlı olacaktır. Elimizdeki veriler içinde bulunan ve sonuca hiçbir katkısı bulunmayan hastaya ait "takma ad" sütunu veri setinden çıkarılmıştır. Benzer şekilde, nodülün "sınır düzenliliği" özniteliği tüm hastalar için aynı sonucu ihtiva ettiğinden ve etkisiz eleman olarak sonuca hiçbir katkı sağlamayacağından, veri setinden çıkarılmıştır. "Nodül sayısı" özniteliğinde ise yeterince veri bulunmadığından, veri setinden çıkarılmıştır. Diğer tüm öznitelik sütunları içinde, toplamda üç adet eksik veri ortalama (*mean*) kullanılarak tamamlanmıştır.

2.1.2. Veri dönüştürme (Data transformation)

Makine öğrenmesi modelleri seçilmeden önce, eldeki veri hazır ve analiz edilebilir durumda olmalıdır. Veri seti oluşturulduktan sonra, orijinal halleri ile kullanılmaları ve Makine öğrenmesi modellerinde denenmeleri sonuçlar açısından çeşitli sorunlara neden olabilir. Veri setlerinde çok fazla veri tipi ve bu veri tiplerinin çeşitli ve farklı büyüklüklerde boyutları mevcuttur ve bu durum sorunların kaynağı haline gelebilir. Elde edilmek istenen doğruluğun en iyi oranda olması amacıyla, modelleme çalışmalarına geçmeden önce, veriler arası veri tipi uyumlarının, dengelerinin ve seviyelerinin ayarlanmaları gerekmektedir.

Veri setleri, bir bağımlı ve en az bir bağımsız değişkenden oluşmalıdır. Bağımsız değişkenlerin çok fazla oluşu, seçilen algoritma ve modellerin yavaş çalışmasına ve bu aşamada yanlış sonuçlar üretmesine neden olabilir. Bu nedenle, seçilecek öznitelik sayısının makul düzeyde olması ve bağımlı değişkeni %100'e olabildiğince yakın seviyede tahmin etmesi en önemli amaçtır. Bu nedenle, modeller üzerinde uygulamalara başlamadan önce, uygun öznitelikler seçilmelidirler.

Verilerin, metin ya da sayısal tiplerde olmasına bağlı olarak farklı algoritmalardan farklı sonuçlar elde edilmesi olasıdır ve bu durum sonucu doğrudan etkiler. Aynı olmayan ölçeklendirmeler, analizlere aynı katkıyı sağlayamazlar. Verilerin sayısallaştırılmaları sonrası yapılması gereken işlemlerden bir diğeri ise, veriler arası korelasyonların iyileştirilmesi için, normalleştirme ya da standartlaştırma adı verilen işlemlerden geçirilerek, belli seviyelerde nicelik dengelerinin oluşturulmasıdır. Veri dönüştürme işlemlerinden biri olan Normalleştirme (*Normalization*), uygun metotlar ile farklı boyutlardaki verilerin en uygun aralıklarda tekrar boyutlandırılmalarını sağlar. Böylelikle, veriler arası korelasyonların bulunması ve yorumlanmaları daha verimli hale gelir. Bu nedenle, eldeki veri setine çeşitli normalleştirme işlemleri uyguladık. Ayrıca, veri seti içindeki sütunların fazlalığı, modellerin çalışma zamanını uzatacağı ve sonuca olumsuz etki yapacağı için, en iyi öznitelik sayılarını en iyi sonuçları verebilecek şekilde seçerek bağımsız değişkenleri olabildiğince azalttık.

2.1.3. Veri azaltma ve veri alt kümesi Oluşturma
(Data reduction and data subset generation)

Veri setleri oluşturulurken mümkün olan en az öznitelik sayısı ihtiva eden setler oluşturmak, ilk ve en önemli adımlardan biri olmalıdır. Böylelikle, model seçimimiz ne olursa olsun, modelde en iyi sonucu verecek en az öznitelik seçimi amacımıza yardım edecektir. En az öznitelik seçimi yapılırken, öznitelik sayısı azaltma işlemi belli kurallara göre yapılmalıdır. Bazı durumlarda ise sayı azaltma yerine öznitelik türetme yoluna gidilmelidir. Örneğin, veri özniteliklerimizden biri olan USG LAP dört adet çok terimli seçimden (polinominal: 0,1,2,3) müteşekkil veri tipine sahiptir. Bu veri tipini İkili (0,1) hale dönüştürmek için ilgili özniteliğe bağlı iki adet daha öznitelik türetilmelidir. Ancak burada dikkat edilmesi gereken işlem, öznitelik artırımı yapıldıktan sonra öznitelik azaltmanın, gerekli ise, öznitelik seçim algoritmalarının kullanımı öncesinde yapılması uygun olacaktır.

Elimizde bulunan tiroit kanser hastası veri seti içerisinde var olan çok terimli veriler, ikili veri tipine çevrilmiştir. Ayrıca boyut olarak büyük olan tüm diğer ondalık veriler ise Normalleştirme işlemleri uygulanarak boyutsal küçültmeye gidilmiştir. Boyutsal küçültme, korelasyon sonuçlarının yorumlanabilir düzeye gelmesini sağlamak için gereklidir. Veri azaltma, veri artırma, boyutsal küçültme türünden işlemlerden sonra öznitelik seçim işlemlerine geçilmiştir.

Öznitelik seçimi (Feature selection), kullanılacak model içinde tüm veri setini en iyi temsil edecek ve en iyi sonucu verecek veri alt setinin seçilmesi için atılacak ilk adımdır. Bu adımda, sonucu en iyi verecek öznitelik kombinasyonlarının bulunması amaçlanmıştır. Bu amaca yönelik geliştirilmiş çeşitli algoritmalar mevcuttur. Bu algoritmalar üzerinden, öznitelikler bağımlı değişkeni etkileme oranında sıralanırlar. Bu sıralama kullanılan testlerin hangisi olduğuna bağlı olarak, sayısal açıdan büyükten küçüğe ya da küçükten büyüğe doğru sıralanırlar.

Elimizdeki veri seti, çeşitli normalleştirme, ölçeklendirme ve sayısallaştırma yöntemleri uygulandıktan sonra, bir kısmı kategorik ve bir kısmı ondalık olan toplamda 24 bağımsız ve 1 adet bağımlı değişkenden oluşmaktadır. Girdi verileri, bazı sütunları ikili veri tipine dönüştürdüğümüzden bağımsız değişken sütun sayısı 24'ten 28'e çıkmıştır. Bağımlı değişkeni etkileyen bağımsız değişkenler (öznitelikler) arasından en iyi öznitelikleri çıkarmak için üç temel yöntem kullanılmaktadır. Bu yöntemler:

Sarmalayıcı yöntem (Wrapper method): Öznitelik kümesi, kullanılan sınıflandırıcının, çeşitli iterasyonlarla ortaya koyduğu en iyi sonuca bağlı olarak seçilirler. Başka bir deyişle Özniteliğin seçilip seçilmeyeceği, iterasyonlar sonrası bağımsız değişkenin sonuca vereceği olumlu ya da olumsuz katkıya bağlıdır. Ardışık ileri yönlü seçim yöntemi ve Ardışık geri beslemeli eleme, Detaylı özellik seçimi (Exhaustive feature selection), Özyinelemeli özellik eleme (Recursive feature elimination) yöntemleri bu kategoriye giren ve sıklıkla kullanılan öznitelik seçim yöntemlerindendir.

Filtreleme yöntemi (Filter method): Filtreleme yaklaşımı, herhangi bir sınıflandırıcı yöntemi önceden belirtilmeksizin veri seti içindeki bağımlı ve bağımsız veriler arasında bağımlılık, bilgi, tutarlılık yönünden ilişkileri göz önüne alan algoritmaları kullanır. Korelasyon temelli ve yardım yaklaşımlı öznitelik seçimi, Bilgi kazanımlı (IG-

information gain), Ki-kare (Chi-square) test, Balıkçı puanlaması (fishers score), Korelasyon katsayısı (Correlation Coefficient), Varyans eşiği (Varience thrashold), Ortalama mutlak fark (Mean absolute difference), Dağılım oranı (Dispersion ratio) gibi algoritmalar filtreleme yaklaşımı olarak sıklıkla kullanılmaktadır.

Gömülü yöntem (Embedded method): Sarmalayıcı ve filtreleme yaklaşımlarının her ikisini kullanan ve her ikisinin avantajlı yanlarını da sonuçlara yansıtmak isteyen bir yaklaşımdır. Öznitelik seçme yöntemlerinden olan Karar ağaçları ve Ölçeklendirilmiş Naif bayes (Naive bayes) yaklaşımı, bu kategoride sıklıkla kullanılan yöntemlerdendir.

Bu üç yaklaşımla yapılan Öznitelik seçimleri, metin madenciliğinde [17-18], kanser tanısında [19], sahtekarlık tespitinde [20], kredi puanlamasında [21], müşteri kazanma/kaybetme nedenleri analizinde [22], güvensiz eposta tespitinde [23] kullanılmışlardır.

Veri setimiz kullanılarak uygun testler ile yapılan sıralamada özniteliklerden ilk 6 adedi seçilmiştir. 28 değişken arasından 6 adetinin seçilmiş olması test kurallarına uygunluğu nedeniyledir. Örneğin, bağımsız değişkenler test sonucunun p-değeri 0-1 arasında sıfıra yakınsadığı oranda önem kazanmış ve seçilecek en önemli öznitelikler arasında yerini almıştır. Altı adetten daha fazla değişkenin- uygulanan testlere göre- sonuca olumlu ve yeterli bir katkı sağlamayacağı sonucuna ulaşılmıştır. Bu testler Excel'e entegre edilen AnalyticSolver eklentisi kullanılarak yapılmıştır. Kullanılan testler ve bu testlere ait kurallar Tablo 4'te gösterilmiştir. Veri setine uygulanan test sonuç tablosu Tablo 5'te gösterilmiştir. Tablo 4'te gösterilen çeşitli testler uygulandıktan sonra elde edilen en önemli öznitelikler, üçüncü bölümde Tablo 10'de üçüncü sütun altında gösterilmiştir.

Tablo 4. Uygulanan testler ve sıralama kuralları
(Tests applied and their sequencing rules.)

Testler	Kurallar
Ki-kare İstatistik (Chi2: stat), Kramer değeri (Cramers V), Müşterek malumat (Mutual Information), Arttırım oranı (Gain ratio), F: İstatistik (F: Stat), Fisher puanı (Fisher Score), Welch:İstatistik (Welch: Stat), Spearman:İstatistik (Spearman: Stat), Kendall : İstatistik (Kendall: Stat)	Testlerin kuralları gereği ortaya çıkan sonuçlar yukarıdan aşağıya büyükten küçüğe doğru sıralanmıştır.
Ki-kare:p-değeri (Chi2: p-Value), Welch:p-değeri (Welch: p-Value), Gini İndeksi (Gini Index), F:p-değeri (F: p-Value), Pearson:p-değeri(Pearson: p-Value), Spearman:p-değeri(Spearman: p-Value), Kendall:p-değeri(Kendall: p-Value)	Testlerin kuralları gereği ortaya çıkan sonuçlar yukarıdan aşağıya küçükten büyüğe doğru sıralanmıştır.

3. Lojistik Regresyon Modeli (Logistic Regression Model)

Eldeki veri seti için sonuca götürecek en iyi makine öğrenmesi modelinin ne olduğu ile ilgili net bir cevap yoktur. Ancak veri çeşitliliğinin uygun veri tiplerine çevrilmeleri, veri setlerinin hazırlanması, verilerin ölçeklendirilmeleri, özniteliklerin seçimi ve Makine öğrenmesi teknikleri/modelleri ile defalarca denenmesi yukarıda sorulan soruya tatmin edici bir cevap sağlayabilir. Bu cevap, verilerin elde edilmesinden varılan sonuca kadar takip edilen her bir basamakta bulunan çeşitli testlere, algoritmalara ve kullanılan tekniklere bağlıdır. Veri setinin hazır hale getirilmesi, hangi makine öğrenmesi tekniği kullanılırsa kullanılsın sonucu kesinlikle etkileyecektir. Veri setlerinde elde edilmek istenen sonuç (bağımlı değişken) evet/ hayır, kanserli/ kansersiz, 1 ya da 0 gibi ikili veri tipi değişkenlerden oluşuyorsa, analizlerde kullanılacak algoritmaların

Tablo 5. Öznitelik seçim testleri ve sıralama (Feature selection tests and their rankings)

Öznitelik Seçimi: İstatistiksel					
Değişken (Variable)	Ki-kare (Chi2)	p- değeri (p-value)	Cramer değeri (Cramers value)	Müşterek Malumat (Mutual information)	Artış oranı (Gain ratio)
Sitoloji_sonucu_iyi huylu	65,5900	5,55174E-16	0,529433184	0,209117541	0,213642934
USG Lap_mevcut	39,6385	3,05594E-10	0,411576781	0,125465891	0,213770520
Sitoloji_sonucu_kötü huylu	39,1518	3,92083E-10	0,409042528	0,131733398	0,266677578
USG_kötü huylu	29,5356	5,48971E-08	0,355275650	0,098380931	0,233591809
USG_olası_iyi huylu	24,9354	5,92818E-07	0,326438077	0,080519135	0,082261604
Sitoloji_sonucu_olası_kötü huylu	19,8616	8,32554E-06	0,291339470	0,050456966	0,057891672
USG_olası_kötü huylu	16,9218	3,89500E-05	0,268915863	0,051604071	0,056461887
Nodül_ekosu_izoekoik	15,1604	9,87504E-05	0,254535336	0,049291212	0,053931305
Mikrokalsifikasyon	14,8850	0,000114264	0,252213031	0,044832649	0,059900717
Nodül_ekosu_hipoekoik	14,0086	0,000181974	0.244675091	0,044873330	0,047133856
Yaş	13,1995	0,153781590	0,237504633	0,041148317	0,013988721
Nodül özelliği	12,6401	0,000377544	0,232417420	0,042537586	0,055662569
Neutrophil	10,3074	0,244108097	0,209878139	0,035305603	0,015928567
Lenfosit	08,5661	0,478250535	0,191330266	0,027049031	0,009841453
Platelet	08,3090	0,503326869	0,188437631	0,027880116	0,010433369
Şüpheli_nodülün_çapı_1	07,8023	0,005217922	0,182601328	0,025373259	0,031678594
MCHC	07,1342	0,522223805	0,174608397	0,025303551	0,011154391
USG_iyi huylu	06,9476	0,008393041	0,172309958	0,023128914	0,035051724
MCV	06,9196	0,328340738	0,171962810	0,022466779	0,012057412
sT4	06,8483	0,335099712	0,171074420	0,024545133	0,016755967
Şüpheli_nodülün_çapı_0,75	06,3700	0,011506560	0,164991582	0,021148891	0,032535633
sT3	05,5528	0,475098078	0,154045187	0,022549797	0,013473248
Şüpheli_nodülün_çapı_0,25	05,5393	0,018593061	0,153859051	0,015799614	0,019991242
Hgb	05,2240	0,814353103	0,149416011	0,015691293	0,005733729
Mezenter indeksi	05,0623	0,652349985	0,147085452	0,015239251	0,005856772
RBC	04,3765	0,525860507	0,136759531	0,014508053	0,005395722
Tiroit_fonksiyon_normal	03,9767	0,045133501	0,130362898	0,012513552	0,014901741
Şüpheli_nodülün_çapı_0	03,5237	0,050494889	0,122714398	0,010588210	0,017707506

doğrusal yerine doğrusal olmayan yöntemler olması uygun olacaktır. Makine öğrenmesi altında ikili sınıflandırma algoritmaları olarak kullanılan birçok algoritma mevcuttur. Bu algoritmalardan biri de Lojistik regresyon (Logistic Regression) dur.

Lojistik regresyon bir algoritma çeşididir ve doğrusal logaritmaları sınıflandırmaya yarar. Matematiksel temsili 1. Ve 2. Denklemlerde gösterilmiştir [24].

$$P(x) = \frac{exp\,(w\,.x+b)}{1+exp\,(w\,.x+b)} \quad (1)$$

$$P(x) = \frac{1}{1+\exp\,(w\,.x+b)} \quad (2)$$

Bu denklemlerde x rasyonel sayılar kümesine ait olmak kaydıyla, girdi değişkenini temsil eder. *P* ise 0 ya da 1 sonucunu veren çıktı (bağımlı) değişkenidir ve olasılık olarak ifade edilir. *W* kullanılan ağırlık katsayısı iken *b* ofset değeridir ve *w.x* ise matrislerin çarpımıdır. Lojistik regresyon elde ettiği sonuçları, her seferinde iki olasılık değerini karşılaştırarak, *x*i en yüksek olasılık değerine sahip olan kategoriye atar. Şekil 3 grafik olarak lojistik regresyon model yaklaşımını temsil etmektedir.

Sonucu veren bağımlı değişken ikili (binary) ise sonuca hangi ihtimalle varıldığını analiz etmek üzere, Lojistik regresyon en iyi yöntemdir [25].

Lojistik regresyon doğrusal olmayan analizler kullanır. Şekil 3'de gösterildiği üzere verilerimiz doğrusal bir grafik üzerinde iyi bir sınıflamaya dahil değilken, Lojistik regresyon grafiği üzerinde anlamlı bir sınıflamaya tabi tutulabilmektedir. Sonuçlar, doğrusal olmayan bir yol izleyerek 1'e ya da 0'a ne kadar yakınsadığını ortaya koyan ihtimaller üzerinden gösterilmektedir.

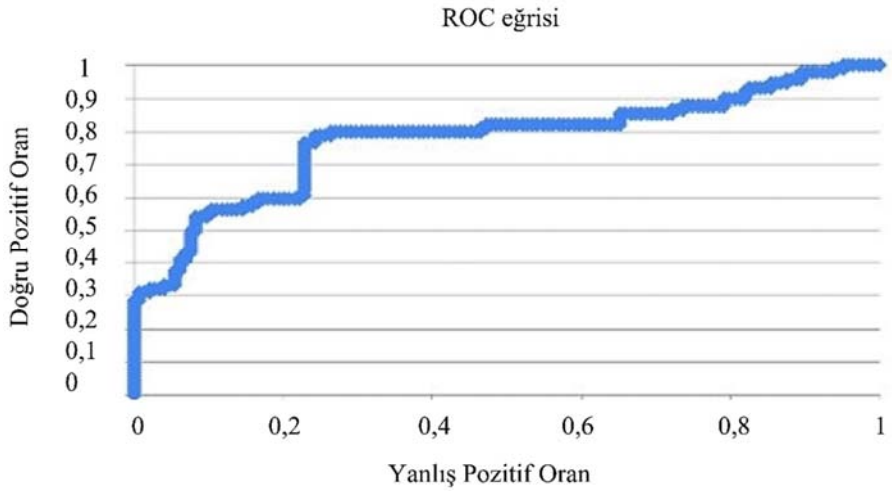
Lojistik regresyonda, yeni bir gözlemin "olumsuz" veya "olumlu" (0 veya 1, bizim veri setimiz için "kansersiz" ve "kanserli" şeklindedir) olduğunu tahmin etmek üzere, en iyi kesme sınırlılık değerini belirlemek için ROC eğrileri kullanılır., Eğer girdi değişkenleri (hasta verileri) seçilen modelde tekrarlar (iterations) sonunda %50 (varsayılan sınırlılık parametresi 0,50 olarak ele alınır) ve üzeri ihtimalle çıktı sonucuna ulaşmışsa kanserli (1), %50nin altında bir ihtimalle sonuca varmışsa kansersiz (0) sınıflamasına dahil olurlar. Lojistik regresyon analiz modeli ile son zamanlarda çeşitli çalışmalar ve denemeler yapılmıştır. Covid-19 hastalığı nedeniyle uygulanan karantina önlemleri sonrası ortaya çıkan riskler üzerinde yapılan çalışmalar ile [26-29], karantina sırasında ortaya çıkan kaygı ve tüm diğer faktörler arasındaki ilişkiyi ortaya çıkarmak için yapılan diğer çalışmalarda [30-33] Lojistik regresyon modelleri kullanılmıştır.

4. Tiroit Kanseri Veri Alt-Kümelerinin Lojistik Regresyon Modelinde Kullanımı
(Use of Thyroid Cancer Data Subsets in the Logistic Regression Model)

Sınıflandırma problemleri için çeşitli makine öğrenmesi modelleri mevcuttur. Denetimli öğrenme yaklaşımlarından biri olan sınıflandırma modellerinde genel itibari ile -istisnalar hariç- modelin performansı, eldeki verinin büyüklüğüne ve kalitesine dayalı olarak genellikle Doğruluk (Accuracy) değeri üzerinden saptanmaktadır. Sınıflandırma algoritmalarında kullanılan veri setinin çıktı değişkenlerinin kategorik olması halinde modelin performans ölçümünde Doğruluk değeri tek başına yeterli gelmez ve hatta yanıltıcı olabilir. Doğruluk ölçümleri yanı sıra Kesinlik (Precision), Duyarlılık (Recall) veya F1 Score ölçümleri birlikte ya da ayrı ayrı olarak daha sağlam bir değerlendirme için gerekebilir. Tüm bu değerlendirmeler modelleme sonrası ortaya çıkan ve Hata matrisi (Confusion matrix) tablosuna yerleştirilen Doğru-pozitif, Yanlış-pozitif, Doğru-negatif, Yanlış-negatif bulguları üzerinden yapılmaktadır.

Hastalık tanısı ile ilgili medikal alanda ikili sınıflandırma problemi söz konusu olduğunda Duyarlılık(recall) ve Kesinlik (presision) yerini aynı formüllerle hesaplanan Duyarlılık (sensitivity) ve Özgüllük (specificity)'e bırakır.

Duyarlılık ve Özgüllük ikili (binary) sınıflandırmada performansın istatistiksel ölçümlerinden ikisidir ve sınıf yaygınlığına bağlı hareket etmez. Duyarlılık doğru-pozitifler arasındaki doğruluğu gösterirken özgüllük ise gerçek negatifler arasındaki doğruluğu ifade eder. Doğru-pozitifler ve doğru-negatifler bu ölçümlerle tamamen ayrı ele alındığından, göreli oranları göz önüne alınmaz. Tıp alanında özellikle tanı gerektiren alanlarda, duyarlılık/özgüllük belirli bir testin özelliklerini taşırlar ve kaç kişinin test edilmekte olunan sonuçlara sahip olduğundan tamamen bağımsızdırlar. Bu, testin istatistiğini zaman ve yer açısından değişmez kılar. Örneğin hastalık durumunun (Kanserli ya da kansersiz) görülme sıklığının %90 olduğu bir popülasyona uygulanan bir test, görülme sıklığının %10 olduğu bir popülasyona uygulandığında aynı duyarlılığa ve özgüllüğe sahip



Şekil 3. %61 sınırlılık düzeyinde ölçümlenmiş olasılık eğrisi gösterimi.
(Probability curve representation measured at the 61% cut-off level.)

olacaktır ve birbirlerinden bağımsızdırlar. Özgüllük genellikle bir tıbbi testi tanımlamanın tercih edilen yoludur. Bir durumun yaygınlığı azaldıkça, sabit bir testin kesinliği azalır, ancak özgüllüğü azalmaz.

Öte yandan Kesinlik, sınıf yaygınlığına bağlıdır. Tahmin edilen pozitifler arasındaki doğruluktur, ancak kaç kişiyi pozitif olarak tahmin ettiğiniz, durumun yaygınlığına bağlı olacaktır. Testi, durumun görülme sıklığının %90 olduğu bir popülasyona uyguladığınızda bir kesinlik değeri elde edersiniz, ancak aynı testi yalnızca %10 insidansa sahip bir popülasyona uyguladığınızda kesinlik çok daha düşük olacaktır. Doğru-pozitif sayısı azaldıkça, pozitif bir testin doğru olma olasılığı da (Kesinlik) azalacaktır.

Kesinlik ve duyarlılık (precision/recall) ikilisi ile duyarlılık (sensitivity =recall)/Özlülük(specificity) ikilisi bir arada uygun ölçütlerdir ve bir diğeri olmadan tek başlarına yeterince ayrıntılıdırlar. Hangi durumlarda hangi ölçüm/ölçümlerin kullanılacağı veri setinin çok iyi analizi ile ortaya çıkarılabilir. Tiroit kanser hastaları için elimizde bulunan veri setini kullanarak yaptığımız sınıflandırmada ve performans değerlendirmesinde Duyarlılık ve Özgüllük ölçütleri iki önemli araç olarak kullanılacaktır.

4.1. Tiroit Kanseri Veri Alt-Kümelerinin İkili Sınıflandırma Algoritmalarında Kullanımı
(Use of Thyroid Cancer Data Subsets in Binary Classification Algorithms)

Lojistik regresyon modeli kullanarak elde ettiğimiz bulguları Hata-matrisi (confussion matrix) üzerinden Duyarlılık ve Özgüllük yorumlamalarına geçmeden önce, ikili sınıflandırma problemlerinde sıklıkla kullanılan bazı sınıflandırma algoritmaları üzerinde farklı hiper-parametreler kullanarak denemeler yaptık. Farklı sınıflandırma algoritmalarında denemeler yapmamızın amacı herhangi bir makine öğrenmesi algoritmasını ele almadan önce veri setimizin farklı sınıflandırma algoritmalarında doğruluk değerleri açısından davranışını saptamaktır. Veri setimiz farklı veri tiplerinden oluşmakta ve çıktı verisi olarak ikili kategorik veriler içermektedir. Veri tipleri göz önünde bulundurulduğunda, ikili sınıflandırma problemlerinde sıklıkla tercih edilen makine öğrenmesi algoritmalarının kullanılması yerinde bir karar olacaktır. Hangi algoritmanın hangi veri seti için iyi sonuçlar vereceği veri setinin büyüklüğüne, veri tipine ve verinin kalitesine bağlıdır ve bu durumlar göz önüne alındığında, Lojistik regresyon algoritmasında olduğu gibi ikili sınıflandırma problemlerinde sıklıkla ve çokça kullanılan K-en yakın komşu (KNN), Destek vektör makineleri (SVM) ve Rastgele orman (Random forest) algoritmalarının deneylerimizde kullanılması uygun görülmüştür. Bu algoritmaların kullanılmasındaki amaç model performanslarının ölçümü değil, verilerin sınıflandırma algoritmalarında davranışlarının tespiti içindir.

K-en yakın komşu sınıflandrma algoritması: Denetimli sınıflandırma algoritmalarından biri olan K-en yakın komşu algoritması doğrusal olmayan veriler için kullanılmaya uygundur. K parametresi verilerin içine dahil edileceği ve oylama (voting) yaklaşımı ile belirlenen sınıflar için kullanılan bir tam sayıdır. Verilerin veri uzayında K içindeki hangi komşuya ait olacağı ile ilgili hesaplamalar çeşitli yaklaşımlarla belirlenebilir. Öklid mesafe (Eucliedan distance) hesaplaması bu yaklaşımlardan biridir.

Destek vektör makineleri (Support Vector Machine) sınıflandırıcı: Daha çok bilinen adı ile Destek vektör makineleri (Support Vector Machine) hızlı-büyük marjlı sınıflandırıcı algoritmasının bir parçasıdır. Destek vektör makine algoritması, eğitim verilerinde bulunan sınıflar arasındaki marjı maksimize eden yüksek bir düzlem bularak verileri sınıflandırmaya çalışır. Bu nedenle, Destek vektör makinesi, büyük marjlı bir sınıflandırıcı örneği olarak ele alınır.

Rastgele Orman (Random Forest) : Tek bir sonuca ulaşmak üzere, birden çok karar ağacının nihai sonuçlarını birleştirir. Rastgele orman algoritması, karar ağaçlarının tahminlerine dayalı olarak çeşitli ağaçlardan elde edilen çıktıların ortalamasını alarak tahminde bulunur. Ağaç sayısını artırmak, sonucun kesinliğini arttırmaktadır.

Tablo 4'te gösterilen testler üzerinden öznitelik seçme işlemleri gerçekleştirilerek, yukarıda kısaca açıklanan algoritmalarda en iyi doğruluk yüzdelerini veren öznitelikler seçilmiştir. Veri setimizde her ne kadar test kurallarımıza uymakta olan belli sayıda (bu örneğimizde 6 adet bulunmuştur) önemli öznitelikler bulunsa da çoğunlukla testler sonrası açığa çıkan en önemli girdi değişkenlerinden sayıca daha azının sınıflandırma modellemelerinde daha iyi sonuçlar verebildiğini tespit ettik. Bu nedenle elde edilen altı adet girdi değişkenli veri seti ile birlikte, bu veri setinden türetilmiş 2li, 3lü, 4lü ve 5li veri seti kombinasyonları kullanılmıştır. Bu veri alt-kümeleri yanı sıra veri setimizin veri tiplerine bağlı olarak oluşturulan farklı veri alt kümeleri de yukarıda açıklanan makine öğrenmesi sınıflandırma algoritmalarında doğruluk değerleri açısından deneye ayrıca dahil edilmişlerdir.

Bu denemeler için Google Colab (Python kodlama ile), Excel (Data Mining ve Makine öğrenme algoritmaları eklenti olarak Excel'e dahil edildi.) kullanıldı. Veri setinde, girdi değişkenlerinin çokluğu, veri tiplerinin farklılığı, işlem zamanı (runtime), varsayılan ayarların (default setting) veya parametre ve hiper-parametrelerin farklılığı türünden nedenlerle aynı modeller kullanılsa da farklı platformlarda farklı sonuçlar verebilmektedir. Bu nedenlerle modeller, daha iyi sonuçlar elde etmek için farklı platformlarda denenmiştir. Bununla birlikte seçilen en iyi özniteliklerin farklı sayıda kombinasyonlarının da farklı modellerde farklı sonuçlar verebileceğini bu şekilde görmüş olduk. Lojistik regresyon modeli ile birlikte, makine öğrenmesi modellerinde denenen veri alt setleri üzerinden Doğruluk sonuçları Tablo 6'da bir arada verilmiştir. Tablo 6'da verilen sonuçlar varsayılan parametrelerin yanı sıra farklı parametre ve hiper-parametreler kullanılarak elde edilen en iyi Doğruluk sonuçlarını da göstermektedir. Bu sonuçların elde edilmesinde eğitim ve test verileri %70 e %30 oranı ile kullanılmıştır. Modeller ve veri seti alt kümeleri sayıca çok olduğundan ve burada her biri için detaylı analiz yapılamayacağından Doğruluk (Accuracy) değerleri üzerinden direkt olarak sonuçlandırılmıştır. Çapraz doğrulama (Cross Validation) değerinin 5 olarak belirlendiği denemelerde elde edilen Doğruluk yüzdelerinin ortalamaları alınmıştır.

4.2. Lojistik Regresyon Algoritmasında Veri Alt-Kümelerinin Kullanımı
(Use of Data Subsets in Logistic Regression Algorithm)

Tablo 10'de birinci ve ikinci sütunda gösterilen ölçeklendirme ve öznitelik seçme testleri ile yapılan işlemler gerçekleştirilmeden önce eldeki tiroit kanseri veri seti tamamı ile sayısallaştırıldı. Sayısallaşma sonrası veriler arası veri-tipi uyumu sağlamak üzere, ondalık dışındaki kesikli ve kategorik veriler ikili veri tipine dönüştürülmek üzere öznitelik çoğaltma (one-hot-codding) işlemine tabi tutuldu.

Tablo 1 de içeriği verilen eldeki orijinal veri seti tamsayı, ikili ve ondalık sayılar ile metin türünden farklı veri tiplerinden oluşmaktadır. Veri seti içindeki nümerik olmayan metin tipi tüm veriler sayısallaştırıldıktan sonra, kesikli tam sayılı veriler 0-1'den oluşan ikili verilere çevrilmiştir. Bunun için örneğin USG sütunu için var olan dört ayrı veri kullanılarak (iyi huylu-0, olası iyi huylu-1, kötü huylu-2, olası kötü huylu-3) ikili veri tipine dönüştürmek üzere dört ayrı sütun oluşturulmuş ve ilgili sütun için 1 "var", 0 ise "yok" amacıyla kullanılmıştır. Cinsiyet, postopSonuc (Kanserli-Kansersiz) gibi iki verili sütunlar ikili (0 ya da 1) sayısal veri tipine çevrilirken ondalık sayılar olduğu gibi bırakılmıştır. Ayrıca ondalık sayılar içinde

Tablo 6. Çapraz Doğrulama kullanılarak farklı veri setleri ve farklı parametreler ile ikili sınıflandırma algoritmalarında ortaya çıkan doğruluk değerli deney sonuçları.
(Experiment results with accuracy values in binary classification algorithms with different data sets and different parameters using Cross Validation.)

Çapraz doğrulama=5	Lojistik regresyon		Destek vektör makineleri		K-en yakın komşu		Rastgele orman	
1313	Varsayılan parametrelerle ortalama doğruluk (Accuracy) (%)	Farklı Hiperparametreler ve doğruluklar(%)	Varsayılan parametrelerle ortalama doğruluk	Farklı Hiperparametreler ve doğruluklar(%)	Varsayılan parametrelerle ortalama doğruluk	Farklı Hiperparametreler ve doğruluklar(%)	Varsayılan parametrelerle ortalama doğruluk	Farklı Hiperparametrelerve doğruluk(%)
Sitoloji_sonucu_iyi_huylu USG_LAP_Mevcut	76,57	77,28 C: >1	76,14	78,56 C:1 kernel: sigmoid	68,42	78,11 neighbors:59	76,56	77,72 estimators: 2
İkili (Binary) veri tipine dönüştürülmüş tüm tamsayılı veriler	72,70	77,56 C: 20	70,98	75,08 C:1 kernel: linear	70,11	76,64 neighbors:10	74,80	76,17 estimators:25
Sürekli (Tam sayılı) veri tipine sahip tüm veriler	74,40	76,64 C: 1	73,54	75,83 C:1 kernel: rbf	72,68	76,56 neighbors:5	76,95	75,31 estimators:25
Sayısal değişkenlerin tümü (Veri setinin tamamı)	73,50	75,66 C: 1	61,54	75,68 C:30, kernel: linear	58,99	61,97 neighbors:30	72,69	70,56 estimators:25
Tüm ondalık değişkenler	58,54	60,26 C: 5	61,54	61,96 C:1 kernel: poly	54,73	61,97 neighbors:10	58,98	58,55 estimators:30
(İkili yapıya dönüştürülmemiş) USG USG_LAP Sitoloji_sonucu	77,38	77,38 C: 1	71,85	74,40 C:1 kernel: linear	70,14	73,53 neighbors:10	71,41	70,98 estimators:15
(İkili yapıya dönüştürülmüş) USG_LAP Sitoloji_sonucu_iyi_huylu Sitoloji_sonucu_olası_kötü_huylu	72,26	72,26 C: 1	70,98	76,99 C:1 kernel: sigmoid	71,42	72,25 neighbors:30	70,13	76,58 estimators:1
(İkili yapıya dönüştürülmüş) USG_olası_iyi_huylu USG_LAP_mevcut Sitoloji_sonucu_iyi_huylu Sitoloji_sonucu_olası_kötü_huylu	75,25	75,25 C: 1	70,15	76,97 C:15 kernel: sigmoid	76,51	76,51 neighbors:5	71,00	76,11 estimators:1
(İkili yapıya dönüştürülmüş) USG_olası_iyi_huylu USG_olası_kötü_huylu USG_LAP_mevcut Sitoloji_sonucu_iyi_huylu Sitoloji_sonucu_olası_kötü_huylu	74,40	74,40 C: 1	70,56	76,09 C:1 kernel: sigmoid	70,95	74,81 neighbors:10	70,13	70,98 estimators:15

virgüllü ayraç yerine nokta konularak veriler kullanılacak algoritma için hazır hale getirilmiştir.

Ölçeklendirme sağlamak üzere ikili (0,1) veriler dışındaki diğer tüm sürekli ve ondalık veriler, Mutlak maksimum ölçeklendirme (Absolute maximum scaling), Min-maks ölçeklendirme (Min-max scaling), Ortalama (mean) kullanarak Normalleştirme (Normalization), Güçlü ölçeklendirme (Robust Scaling), Standartlaştırma (Standardization) yöntemleri ile ölçeklendirildiler. Dört farklı ölçeklendirme işlemiyle oluşan dört farklı veri seti, alt veri setleri oluşturmak üzere Tablo 4'te gösterilen testlere tabi tutulmuşlardır. Ortaya çıkan en uygun veri alt setleri içindeki en uygun öznitelikler, Lojistik regresyon modelinde 6'lı, 5'li, 4'lü, 3'lü ve 2'li kombinasyonlar halinde işleme tabi tutuldular.

Sitoloji_sonucu_iyi_huylu ve USG_LAP_Mevcut girdi değişkenlerinden oluşan veri seti kullanılarak Excel'de (Data Mining ve Makine öğrenme algoritmaları eklenti olarak Excel'e dahil edildi.) çalıştırılan Lojistik regresyon algoritması tahmin yüzdeleri oluşturmuştur. Bu tahmin yüzdeleri ile elde edilen sonuçların Olasılıklı regresyon değerleri ve bu değerlere karşılık gelen doğru tahminlerin bir kısmı Tablo 7'de kesit olarak gösterilmiştir.

Lojistik regresyon algoritması (Python yazılımına ait sklearn kütüphanesi üzerinden) varsayılan parametreler (penalty=l2 tol=0.0001, C=1.0, intercept_scaling=1, class_weight=None, random_state=None, solver=lbfgs, max_iter=100, multi_class=auto) ve 0,50 ve 0,61 sınırlılık (cut-off) ile ayrı ayrı çalıştırılması ile elde edilen bulgular iki ayrı hata matrisinde ayrı ayrı gösterilmiştir. 0,50 sınırlılık ile elde edilen Hata matrisi Tablo 8'da verilirken 0,61 sınırlılık ile elde edilen hata matrisi Tablo 9'da verilmiştir.

Lojistik regresyon algoritmasında 0,50 sınırlılık değeri üzerinden hesaplanan ve Tablo 8'da gösterilen sonuçlara göre performans ölçüm değerleri;

- Doğruluk oranı (Accuracy rate) = (TP+TN) / (TP+TN+FP+FN) = (69 + 111) / 234 = **0.7692**
- Kesinlik veya Pozitif Öngörü Değeri (Precision or the Positive Predictive Value) =TP/(TP+FP) = 69 / 102 = **0,6765**
- Anımsama, Duyarlılık veya Gerçek Pozitif Oran (Recall or Sensitivity or True Positive Rates)=TP/(TP+FN)=69 / 90=**0,7667**
- Spesifiklik/Özgüllük veya gerçek negatif oran (Specificity or True Negative rate) = TN/(TN+FP) = 111/144=**0,7708**

olarak hesaplanmıştır.

Benzer şekilde Lojistik regresyon algoritmasında 0,61 sınırlılık değeri üzerinden hesaplanan ve Tablo 9'da gösterilen sonuçlara göre performans ölçüm değerleri;

- Doğruluk oranı (Accuracy rate) = (TP+TN) / (TP+TN+FP+FN) = (27 + 143) / 234 = **0,7265**

Tablo 7. Olasılıklı regresyon tablosu (Probit regression table)

A	*B*	*C*	*D*	*E*	*F*
Bağımlı değişken (Post-op)	*p-tahmin (p-Pred)*	*Başarılı tahmin (Suc-Pred)*	*Başarısız tahmin (Fail-Pred)*	*LL*	*% Doğruluk (%Correct)*
1	0,152314935	0,152314935	0,84768506	-1,8818	0
0	0,152314935	0,152314935	0,84768506	-0,16525	100
0	0,152314935	0,152314935	0,84768506	-0,16525	100
0	0,594414738	0,594414738	0,40558526	-0,90242	100
0	0,152314935	0,152314935	0,84768506	-0,16525	100
0	0,152314935	0,152314935	0,84768506	-0,16525	100
0	0,936449935	0,936449935	0,06355007	-2,75593	0
1	0,152314935	0,152314935	0,84768506	-1,8818	0
0	0,152314935	0,152314935	0,84768506	-0,16525	100
1	0,594414738	0,594414738	0,40558526	-0,52018	0
1	0,936449935	0,936449935	0,06355007	-0,06566	100
0	0,594414738	0,594414738	0,40558526	-0,90242	100
1	0,594414738	0,594414738	0,40558526	-0,52018	0
0	0,152314935	0,152314935	0,84768506	-0,16525	100
0	0,152314935	0,152314935	0,84768506	-0,16525	100
0	0,152314935	0,152314935	0,84768506	-0,16525	100
0	0,152314935	0,152314935	0,84768506	-0,16525	100
0	0,152314935	0,152314935	0,84768506	-0,16525	100
0	0,594414738	0,594414738	0,40558526	-0,90242	100
0	0,152314935	0,152314935	0,84768506	-0,16525	100
0	0,152314935	0,152314935	0,84768506	-0,16525	100
1	0,602632139	0,602632139	0,39736786	-0,50645	0
…„	……„	……	……	,……,	…„…

Tablo 8. 0,50 sınırlılık altında Karışıklık matrisi tablosu (Confusion matrix table under 0.50 cut-off)

	Başarılı-Gerçek (Suc-Obs)	Başarısız-Gerçek (Fail-Obs)	Toplamlar
Başarılı-Tahmin (Suc-Pred)	69 (TP)	33 (FP)	102 (PP)
Başarısız-Tahmin (Fail-Pred)	21 (FN)	111 (TN)	132 (PN)
	90 (OP)	144 (ON)	234 (TOPLAM)
	0.766667	0.770833	0.769231
Cutoff : 0.50	*PP = tahmini pozitif = TP + FP | PN = tahmini negatif = FN + TN*		
TP:Doğru pozitif | FP:Yanlış pozitif | FN:Yanlış negatif | TN:Doğru negatif | OP=gözlemlenen pozitif (TP+FN) | ON: FP+TN			

Tablo 9. 0,61 sınırlılık altında Karışıklık matrisi tablosu (Confusion matrix table under 0.61 cut-off)

	Başarılı-Gerçek (Suc-Obs)	Başarısız-Gerçek (Fail-Obs)	Toplamlar
Başarılı-Tahmin (Suc-Pred)	27 (TP)	1 (FP)	28 (PP)
Başarısız-Tahmin (Fail-Pred)	63 (FN)	143 (TN)	206 (PN)
	90 (OP)	144 (ON)	234 (TOPLAM)
	0.30	0.993055556	0.726495726
Cutoff : 0.61	*PP = tahmini pozitif = TP + FP | PN = tahmini negatif = FN + TN*		
TP:Doğru pozitif | FP:Yanlış pozitif | FN:Yanlış negatif | TN:Doğru negatif | OP=gözlemlenen pozitif (TP+FN) | ON: FP+TN			

- Kesinlik veya Pozitif Öngörü Değeri (Precision or the Positive Predictive Value) =TP/(TP+FP) = 27 / 28 = **0,9643**
- Anımsama, Duyarlılık veya Gerçek Pozitif Oran (Recall or Sensitivity or True Positive Rates) =TP/(TP+FN) = 27/90=**0,30**
- Spesifiklik/Özgüllük veya gerçek negatif oran (Specificity or True Negative rate) = TN/(TN+FP) = 143/144= **0,9931**

olarak hesaplanmıştır.

Her iki karışıklık matris tablosuna ve elde edilen yukarıdaki sonuçlara göre;

Doğruluk oranı doğru tahmin edilen pozitif sonuçlar ile doğru tahmin edilen negatif sonuçların toplamının tüm sonuçların toplamına bölünmesi ile elde edilen bir orandır. Birçok sınıflandırma modellerinde performans ölçüsü olarak kullanılsa da tıbbi tanı problemlerinde kullanılan sınıflandırma modellerinin performans ölçümlerinde için tek başına yeterli olmayabilir. Her iki sınırlılık oranına göre elde dilen doğruluk oranı %70 üzerindedir ve bu oranlar çok iyi olmasa da her iki durum için de kötü sonuç sayılmazlar. Kesinlik veya pozitif öngörü değeri, yanlış pozitif oluşumlarının kabul edilemez olduğu durumlarda kullanılır. Doğru pozitiflerin artması kesinlik değerini arttıracaktır. Sınırlılık değerinin 0,50 olduğu durumda tahmin edilen toplam pozitiflerin içinde yanlış pozitiflerin çokluğu sonucu 0,6765 ile düşük çıkarmıştır. Sınırlılık düzeyinin 0,61 olduğu durumda ise doğru tahmin edilen ve gerçekte de doğru olan pozitif sayısının yanlış pozitif sayısına oranla çok fazla olması kesinlik sonucunu %96,46 olarak açığa çıkarmıştır.

Duyarlılık veya Gerçek Pozitif Oran ise gerçek pozitiflerin verilerdeki toplam (gerçek) pozitiflere oranıdır. Yanlış negatiflerin oluşmasının kabul edilemez olduğu durumlarda kullanılır. Tablo 8'da yanlış negatifler Tablo 9'daki yanlış negatiflere oranla daha azdır ve duyarlılık yüzdesi daha fazladır. Yanlış negatiflerin yanlış pozitiflerden daha az olması beklenir. 0,50 sınırlılık oranıyla Tablo 8'da duyarlılık oranı daha iyi bir sonuç ortaya koymuştur. Gerçek kanserliler içinden doğru kanser tanısı koyma oranı Tablo 8'da

Tablo 10. Farklı öznitelik kümeleri ile elde edilen sonuçlar (Accuracy results obtained with different feature sets.)

Ölçeklendirme	Kullanılan testler	En önemli öznitelikler	En iyi öznitelikler	Özgüllük (Specificity) (%61 sınırlılık)
Absulute Max Min-Max Normalization by MEAN Robust Scaling Standardization	Chi2: Stat, p-Value, Cramers V Welch: p-Value, F: Stat, p-Value, Fisher Score Gini Index, Pearson: Stat, p-Value Kendall: Stat, p-Value Spearman: Stat, p-Value	Sitoloji_sonucu_iyi huylu USG LAP_Mevcut Sitoloji_sonucu_kötü huylu USG_kötü huylu USG_olası_iyi huylu Sitoloji_sonucu_olası_kötü huylu	Sitoloji_sonucu_iyi huylu USG_LAP_Mevcut	%99,31
Absulute Max Min-Max Normalization by MEAN Robust Scaling Standardization	Mutual Information	Sitoloji_sonucu_iyi huylu Sitoloji_sonucu_kötü huylu USG LAP_Mevcut USG_kötü huylu USG_olası_iyi huylu Sitoloji_sonucu_olası_kötü huylu	Sitoloji_sonucu_iyi huylu Sitoloji_sonucu_kötü huylu	%99,31
Absulute Max Min-Max Normalization using MEAN Robust Scaling	Gain Ratio	Sitoloji_sonucu_kötü huylu USG_kötü huylu USG LAP_Mevcut Sitoloji_sonucu_iyi huylu USG_olası_iyi huylu Sitoloji_sonucu_olası_kötü huylu	Sitoloji_sonucu_kötü huylu USG_kötü huylu	%98,61
Robust Scaling Standardization	Welch:Stat	Sitoloji_sonucu_iyi huylu USG_olası_iyi huylu Nodül ekosu_izoekoik Şüpheli Nodülün Çapı_1 USG_iyi huylu Şüpheli Nodülün Çapı_0.75	Sitoloji_sonucu_iyi huylu USG_olası_iyi huylu Nodül ekosu_izoekoik	%85,42

%76,67 iken Tablo 9'daki verilere göre bu oran %30 olarak çok düşük gerçekleşmiştir. Bunun nedeni gerçek kanserliler içinden yanlış kanserli tanısı koyma sayısının çok fazla olmasıdır (FN=63).

Spesifiklik/Özgüllük veya gerçek negatif oran verilerdeki gerçek negatiflerin toplam negatiflere oranı olarak ifade edilir. Spesifiklik, algoritma tarafından gerçekte sağlıklı olan herkese doğru negatif etiketinin verilmesidir. Bizim tiroit kanseri veri setimizde sıfır (0) ile gösterilen sonuçlar patolojide negatif çıkan (gerçekte kanser olmayan) ancak yanlışlıkla ameliyat edilen hastaları temsil etmektedir. Bu durumda kullandığımız algoritmanın tüm bu sıfırlar içinden (kanser olmayanların tamamı içinden) olabildiğince çoğunluğu negatif (kanser değildir) olarak işaretlemektir. Bu durumda spesifikliğin yüksek oranda olması kanserli olmayanları olabildiğince yüksek oranda açığa çıkarmamız anlamına gelmektedir. Sınırlılığın 0,61 olduğu Tablo 9 verilerine göre 144 adet kansersiz hasta içinde kansersiz tanısı koyma sayısı 143 iken yanlışlıkla Kansersiz bir kişiye kanserlidir teşhisi koyulmuştur. Spesifiklik ölçüsüne göre algoritma, gerçekten kanser olmayanlara kansersiz oldukları tanısını koyarak %99,31 oranını elde etmiştir. Uygulama sonucu ortaya çıkan en yüksek özgüllük (Specificity) değeri 0,61 sınırlılık ile elde edilmiş ve Tablo 10'da son sütunda gösterilmiştir.

5. Sonuçlar (Conclusions)

Makine öğrenmesi algoritmaları daha önce sonuçları uzmanlar tarafından ispatlanmış kalıpları öğrenerek örüntüler oluşturabilirler. Elde edilen bu örüntüler üzerinden, daha sonra gelen farklı hasta verileri işleme sokulabilir, tanı sonuçları hızla elde edilebilir ve tahlil/tetkik sayıları azaltılarak aynı zamanda maliyet düşürülebilir.

Yöntemler, teknikler ve algoritmalar, gerçek hayat senaryolarına en yakın olacak şekilde seçilmelidir. Algoritmalarla elde edilen sonuçlar, gerçek hayatta elde edilmiş doğru sonuçlara olabildiğince yakınsamalı ve bu nedenle klinik analizler ile istatistiki analizler arası farklar en aza indirilmelidir.

Lojistik regresyon modeli kullanımında Excele eklenen xlMiner ve Xrealstats, Analytic Solver eklentileri üzerinden Real Statistics menüsü oluşturularak öznitelik seçimi için Sıralı ileri adımlı seçim (Squential Forward Selection) yöntemi kullanılmıştır. Lojistik regresyon sınıflandırmasında deneyler yapılmadan önce veri ön işleme yöntemleri ile hazır hale getirilen veri setimiz ayrıca farklı veri tiplerinden oluşan farklı alt veri kümelerine ayrılmıştır. Tablo 6'da Kullanılan girdi değişkenleri sütunu altında gösterilen farklı veri alt kümeleri K-en yakın komşu, Destek vektör makineleri ve Rastgele orman sınıflandırma algoritmalarında denenmiştir. Lojistik regresyon ile benzer sınıflandırma özellikleri taşıyan bu algoritmalarla farklı veri tiplerinden oluşan farklı veri setlerinin Doğruluk (Accuracy) oranı üzerinden davranışlarını irdeledik. Sonuçları Tablo 6'da gösterilen doğruluk sonuçlarına göre ondalık veriler tüm parametreler altında daha düşük sonuçlar verirken diğer veri tiplerinden oluşan veri alt kümeleri farklı parametreler ve hiper-parametreler altında görece daha iyi sonuçlar vermiştir.

Bununla birlikte en uygun Makine öğrenmesi tekniği seçimi için literatür taraması yapıldı ve ikili sonuç ihtiva eden veri setlerinde Lojistik regresyon modeli kullanımının elimizdeki veri setine uygun olduğu kararlaştırıldı. Bu modele karar verirken, bağımlı değişkenimizin (Kanserli, Kansersiz- 0,1) doğrusal çözümlemeler yerine doğrusal olmayan çözümlemelere (Lojistik regresyon) uygun

olduğu göz önüne alınmıştır. Lojistik regresyon ile elde edilen sonuçları Karışıklık matris tablosu (Confusion matrix table) üzerinden yorumlamadan önce, belirtmeliyiz ki bağımlı değişkenimiz olan postopSonucu, ameliyat gerçekleştikten sonra alınan parçaların patoloji sonuçlarıdır. Buna göre patoloji sonucu (Post-op) kansersiz (0) olarak ortaya çıkanların ameliyat olmaması gereken hastalar olduğu, dolayısı ile verilen ameliyat kararının yersiz olduğu ortaya çıkmaktadır. Bu yaklaşımla bakıldığında Karışıklık matrisinde kansersiz (0) veriler spesifik (Specificity) olarak ele alınmalı ve bu veriler üzerinden yorumlamalar yapılmalıdır.

Sonucu Kanserli ya da Kansersiz olarak tayin eden Lojistik regresyon, tayinini olasılıklar üzerinden yapmaktadır. Olasılığı yüzde olarak açığa çıkaran bu model, kanserli ya da kansersiz vakaların hangi yüzdelik dilimine girdiğine karar verir. Bu kararın sonucu, bizim tayin edeceğimiz yüzdeliğin sınırına bağlı olarak değişecektir. Yüzdelik sınırının iyi tayini ve öznitelikelerin uygun seçilmiş olması, yüksek sonuçları elde edilmesini sağlayacaktır. Lojistik regresyon sonucu elde edilen ihtimal belirlenen sınırlılığın (varsayılan sınırlılık parametresi 0,50) altında ise kansersiz, belirlenen sınırlılığın üzerinde ise Kanserli sonucu verecektir.

Tiroit kanser vakaları özelinden hareketle, burada amaç gereksiz ameliyatların önüne geçmek olacağından bu modelde kansersiz (kanser olmama) sonuçları üzerinden gidilmiştir. Bu nedenle, her bir veri setini ve farklı sayıda öznitelik kombinasyonları kullanarak yüzdelik sınırını %50 sınırından başlattık ve birer yükselterek her seferinde modelimizi çalıştırdık ve sonuçlar elde ettik. Sınırlılığın 0,50 ve 0,61 olduğunda elde edilen sonuçlar Tablo 8 ve Tablo 9'da gösterilmiştir. %61 sınırlılık (cut-off) ile elde edilen ihtimal sonuçları kesit olarak Tablo 8'de verilmiştir. Tablo 8'deki verilere göre F sütunundaki değerler elde edilirken, A sütunundaki bağımlı değişken kanserli (1) ve B sütunu değeri 0,61 sınırlılık oranının altında ise F sütununda karşılık gelen %Doğru (%Correct) değeri %0'dır ve tahmin doğru çıkmamış demektir, B sütunu değeri 0,61 sınırlılık oranının üstünde ise %Doğru (%Correct) değeri %100'dür ve tahmin doğru çıkmış demektir. Benzer şekilde A sütunundaki bağımlı değişken kansersiz (0) ve D sütunu 0,61 sınırlılık oranının altında ise F sütununda karşılık gelen %Doğru (%Correct) değeri %0 ve tahmin doğru çıkmamış, D sütunu 0,61 sınırlılık oranının üstünde ise %Doğru(%Correct) değeri %100 ve tahmin doğru çıkmış demektir. Tüm satırlara uygulanan aynı kurallar sonrası ortaya çıkan sonuçlar Tablo 9'da Karışıklık matrisi tablosunda (Confusion matrix table) gösterilmiştir.

Sınıflandırma model sonuçları üzerinde yapılan değerlendirmelerden biri de Spesifiklik (Specificity) analizidir. Üzerinde çalıştığımız model bağlamında Spesifikliği, belirli bir duruma ait olma anlamında ele aldık. Bu yaklaşımla Spesifikliği kansersiz (0) olma sonuçları üzerinden ele aldık.

Spesifiklik ("gerçek negatif oran" olarak da adlandırılır), doğru bir şekilde tanımlanan negatiflerin (kansersiz-0) oranını ölçer (örn. hastalığa sahip olmadığı doğru şekilde tanımlanan ve gereksiz ameliyat olan sağlıklı insanların yüzdesi).

Buna göre Lojistik regresyon modeli kullanılarak elde edilen 6 adet en iyi öznitelik arasından, deneyerek ortaya çıkardığımız Sitoloji_sonucu_iyi_huylu ve USG LAP_Mevcut öznitelik ikilisi Özgüllülük (Specificity) olarak %99,31 ile en iyi sonucu vermektedir. Bunun anlamı şudur: Elimizdeki tiroit kanser ön teşhisi konulanlar arasından Post-op (Son Operasyon) sonuçlarına göre her 100 kişiden 99,31'i operasyon geçirmiş ancak kanserli olduğu yönünde teşhisinin yanlış konulduğu ortaya çıkmıştır (kansersiz tanısı konulanların gerçekten kanser olmayan tüm kansersizlere oranı). Bu sonuç, ameliyat edilmemesi gereken ancak ameliyat edilen 144 hastanın 143'üne denk gelmektedir. Bunun sonucunda, verilere göre, kanser olma tanısı kullandığımız Lojistik regresyon modeli üzerinden sadece yaklaşık 1 kişide doğru verilmiştir (Kanser ön teşhisi konmuş, ameliyat edilmiş ve sonrasında patoloji ile kanserli olduğu da teyit edilmiştir). Elde edilen sonucun yüksekliği yanında uzman görüşü alınarak seçilen özniteliklerin doğruluğu teyit edilmiştir. Buna göre Sakarya Üniversitesi Araştırma Hastanesi Genel Cerrahi Bölümü hocaları ile yapılan sonuç değerlendirmesinde elde edilen Sitoloji_sonucu_iyi huylu ve USG LAP_Mevcut öznitelikleri, ameliyat öncesi İİA testi yapma konusunda karar almada göz önüne aldıkları girdi değişkenleri olduğu teyit edilmiştir.

Geliştirdiğimiz model tiroit kanseri için kullanılabilir non-invaziv tanı yöntemidir. Bu model ile gereksiz ameliyatların ve tetkiklerin önüne geçilerek sağlıkta maliyet ve komplikasyonlar düşürülebilir. Bu çalışma, tiroit kanseri tanısını öngörmek için uygulanabilir, kolay, hızlı ve ucuz bir yöntem sunmaktadır.

Burada, ayrıca gözden kaçması muhtemel çok önemli bir yaklaşımı belirtmeliyiz. İstatistik ve matematik elbette birçok alanda çözümler sunmaktadır ancak söz konusu tıbbi tahliller, tanı ve tedavi olduğunda, klinik analizler ile istatistiki analizler arasındaki farklılıklar ve benzerlikler iyi kavranmalı ve klinik analizleri istatistiki olanlara kurban vermemelidir. Tüm bu sonuçlara, eldeki verilere uygulanan istatistiki testler üzerinden kullanılan algoritmalarla varıldığı göz önünde bulundurulmalıdır.

Mevcut veri setimizdeki diğer test verilerinin (özellikle kan testlerinden oluşan ondalık veri tipine sahip özniteliklerin) kanser teşhisi kararlarında ne kadar yol gösterici olabileceğini şimdilik, Lojistik regresyon modeli üzerinden gösterebilmiş değiliz. İlgili uzman doktorların özelikle talep ettikleri sonuç, non-invaziv yöntemlerle karar verebilmek için, kan ve hormon testlerinin ne kadar etkili olduğu ile ilgilidir. Bu durum farklı makine öğrenmesi algoritmaları ile test edilerek ayrıca araştırılabilir. Tüm bunlarla birlikte, son yıllarda görüntüleme tekniklerinin modernleşmesi neticesinde elde edilen görüntü verileri üzerinden de Makine öğrenmesi modelleri kullanılabilir. Bu durumda, hastalardan istenen birçok teste gerek kalmaksızın görüntü verileri üzerinden kanser teşhisi konulabilir.

Kaynaklar (References)

1. Niederhuber J.E., Armitage J.O., Doroshow J.H., Kastan M.B., Tepper J.E., Abeloffs Clinical Oncology, 6th edition, Elsevier Publishing, Philadelphia, A.B.D., 2020.
2. National Library of Medicine. MedLinePlus. Thyroid Cancer. https://medlineplus.gov/thyroidcancer.html, Erişim tarihi Temmuz 18, 2023.
3. Ferlay J., Ervik M., Lam F., Global Cancer Observatory: Cancer Today, International Agency for Research on Cancer, Lyon, France, 2020.
4. Lortet T. J., Franceschi S., Dal M.L., Vaccarella S., Thyroid cancer "epidemic" also occurs in low- and middle-income countries, International Journal of Cancer (IJC), 144 (9), 2082-2087, 2019.
5. Li M., Dal Maso L., Vaccarella S., Global trends in thyroid cancer incidence and the impact of overdiagnosis, Lancet Diabetes Endocrinol, 8 (6), 468-470, 2020.
6. Grani G., Sponziello M., Pecce V., Ramundo V., Durante C., Contemporary Thyroid Nodule Evaluation and Management. The Journal of Clinical Endocrinol Metabolism, 105 (9), 2869-2883, 2020.
7. Li M., Brito J.P., Vaccarella S., Long-term declines of thyroid cancer mortality: an international age–period–cohort analysis, Thyroid 30 (6), 838-846, 2020.
8. Grani G., Sponziello M., Pecce V., Ramundo V., Durante C., Contemporary Thyroid Nodule Evaluation and Management, The Journal of Clinical Endocrinol Metabolism, 105 (9), 2869-2883, 2020.
9. Feldkamp J., Führer D., Luster M., Musholt T.J., Spitzweg C, Schott M., Fine Needle Aspiration in the Investigation of Thyroid Nodules, Deutsches Arzteblatt international, 113 (20), 353-362, 2016.

10. Le A.R., Thompson G.W., Hoyt B.J., Thyroid Fine-needle aspiration biopsy: an evaluation of its utility in a community setting, The Journal of Otolaryngol Head Neck Surgery, 44 (1), 12-23, 2015.

11. Fett M.J, Technology, Health and Health Care, 5, Department of Health and Aged Care, Canberra, Australia, 2000.

12. Saluvan M., The Role of Information Systems in Improving the Quality of Health Services, Journal of Health Sciences, 2 (1), 25–39, 2013.

13. Coltin K. L., Using Information Technology to Improve the Quality of Health Care, Health Care Online, 272 (23), 123–158, 1995.

14. Larosa E., Danks D., Impacts on Trust of Healthcare AI Roles for Healthcare AI, AIES '18: Proceedings of the 2018 AAAI/ACM Conference on AI, New Orleans, ABD, 210-2015, 2018.

15. Luca M., Kleinberg J., Mullainathan S., Algorithms Need Managers Too. Harvard Business Review. https://hbr.org/2016/01/algorithms-need-managers-too. Yayın tarihi Ocak-Şubat, 2016. Erişim tarihi Haziran 16, 2020.

16. Hossen S., Big Data And Pattern Recognition, Machine Learning and Big Data Concepts Algorithms Tools and Applications, Dulhare N.U., Scrivener Publishing LLC, 1, 73-103, 2020.

17. Forman G., An Extensive Empirical Study of Feature Selection Metrics for Text Classification, Journal of Machine Learning Research, 3, 1289–1305, 2003.

18. Liu T., Liu S., Chen, Z., Ma W.Y., An Evaluation on Feature Selection for Text Clustering, Proceedings of the Twentieth International Conference on Machine Learning (ICML-2003), Washington D.C, A.B.D., 488-495, 21-24 August, 2003.

19. Bolon V.C., Sanchez N.M., Alonso A.B., Benitez J.M., Herrera F., A review of microarray datasets and applied feature selection methods, Information Sciences, 282, 111–135, 2014.

20. Pouramirarsalani A., Khalilian M., Nikravanshalmani A., Fraud detection in E-banking by using the hybrid feature selection and evolutionary algorithms, IJCSNS, 17(8), 271- 279, 2017.

21. Wang D., Zhang Z., Bai R., Mao Y., A hybrid system with filter approach and multiple population genetic algorithm for feature selection in credit scoring, Journal of Computational and Applied Mathematics, 329, 307-321, 2017.

22. Subramanya K. B., Somani A., Enhanced feature mining and classifier models to predict customer churn for an E-retailer, Cloud Computing Data Science & Engineering Confluence, 7th International Conference, Noida-Hindistan, 531-536, 12-13 January, 2017.

23. Mohamad M., Selamat A., An evaluation on the efficiency of hybrid feature selection in spam email classification, Computer, Communications and Control Technology (I4CT) 2015 International Conference, Kuching-Malezya, 227-231, 21-23 April, 2015.

24. Ananthakumar U., Sarkar R., Application of Logistic Regression in Assessing Stock Performances, IEEE 15th Intl Conf on Dependable, Autonomic and Secure Computing, Pervasive Intelligence and Computing, Big Data Intelligence and Computing and Cyber Science and Technology Congress, Orlando-A.B.D., 1242-1247, 06-10 November, 2017.

25. Joseph S., Munn Brent A., Lanting Steven J., MacDonald Lyndsay E., Somerville Jacquelyn D., Marsh Dianne M., Bryant Bert M., Logistic Regression and Machine Learning Models Cannot Discriminate Between Satisfied and Dissatisfied Total Knee Arthroplasty Patients, The Journal of Arthroplasty, 37, 267-273, 2022.

26. Rossi R., Socci V., Talevi D., Mensi S., Niolu C., Pacitti F., Di Marco A., Rossi A., Siracusano A., Di Lorenzo G., COVID-19 pandemic and lockdown measures impact on mental health among the general population in Italy, Front Psychiatry, 11 (790), 125-132, 2020.

27. Alfawaz H., Yakout S.M., Wani K., Aljumah G.A., Ansari M.G.A., Khattak M.N.K., Hussain S.D., Al-Daghri N.M., Dietary intake and mental health among Saudi adults during COVID-19 lockdown, International Journal of Environmental Research and Public Health, 18 (4), 153-165, 2021.

28. Ting D.S.J., Krause S., Said D.G., Dua H.S., Psychosocial impact of COVID-19 pandemic lockdown on people living with eye diseases in the UK, Eye, 35 (7), 2064–2066, 2021.

29. Jacob L., Smith L., Armstrong N. C., Yakkundi A., Barnett Y., Butler L., Tully M. A., Alcohol use and mental health during COVID-19 lockdown: A cross-sectional study in a sample of UK adults, Drug and alcohol dependence, 219, 186-198, 2021.

30. Fu W., Yan S., Zong Q., Luxford D.A., Song X., Lv Z., Lv C., Mental health of college students during the COVID-19 epidemic in China, The Journal of Affective Disorders, 280, 7–10, 2021.

31. Jiang W., Liu X., Zhang J., Feng Z., Mental health status of Chinese residents during the COVID-19 epidemic, BMC Psychiatry, 20 (1), 1–14, 2020.

32. Liu Z., Liu R., Zhang Y., Zhang R., Liang L., Wang Y., Wei Y., Zhu R., Wang F., Latent class analysis of depression and anxiety among medical students during COVID-19 epidemic, BMC Psychiatry, 21 (1), 498-509, 2021.

33. Liu Y., Chen H., Zhang N., Wang X., Fan Q., Zhang Y., Huang L., Hu B.O., Li M., Anxiety and depression symptoms of medical staff under COVID-19 epidemic in China, The Journal of Affective Disorders, 278, 144–148, 2021.