TİP 2 DİYABET MELLİTUS İLE İLİŞKİLİ RİSK FAKTÖRLERİNİ SAPTAMADA ÇOK DEĞİŞKENLİ İSTATİSTİKSEL YÖNTEMLERİNİN KARŞILAŞTIRILMASI

İpek BALIKÇI ÇİÇEK

BİYOİSTATİSTİK ve TIP BİLİŞİMİ ANABİLİM DALI

Tez Danışmanı
Prof. Dr. Saim YOLOĞLU

Yüksek Lisans Tezi – 2018

T.C.

İNÖNÜ ÜNİVERSİTESİ

SAĞLIK BİLİMLERİ ENSTİTÜSÜ

TİP 2 DİYABET MELLİTUS İLE İLİŞKİLİ RİSK FAKTÖRLERİNİ SAPTAMADA ÇOK DEĞİŞKENLİ İSTATİSTİKSEL YÖNTEMLERİNİN KARŞILAŞTIRILMASI

İpek BALIKÇI ÇİÇEK

Biyoistatistik ve Tıp Bilişimi Anabilim Dalı

Yüksek Lisans Tezi

Tez Danışmanı

Prof. Dr. Saim YOLOĞLU

Bu Araştırma İnönü Üniversitesi Bilimsel Araştırma Projeleri Birimi Tarafından 2016/146Y.Lisans Proje numarası ile desteklenmiştir.

MALATYA

2018

KABUL VE ONAY SAYFASI

İnönü Üniversitesi Sağlık Bilimleri Enstitüsü Biyoistatistik ve Tıp Bilişimi Anabilim Dalı Yüksek Lisans Programı çerçevesinde yürütülmüş olan; **İpek BALIKÇI ÇİÇEK 'in "Tip 2 Diyabet Mellitus İle İlişkili Risk Faktörlerini Saptamada Çok Değişkenli İstatistiksel Yöntemlerinin Karşılaştırılması"** konulu bu çalışması, aşağıdaki jüri tarafından Yüksek Lisans tezi olarak kabul edilmiştir.

Tez Savunma Tarihi: 18/12/2018

Prof. Dr. Saim YOLOĞLU
İnönü Üniversitesi
Tez Danışmanı
Jüri Başkanı

Dr.Öğr.Üyesi Harika Gözde GÖZÜKARA BAĞ
İnönü Üniversitesi
Üye

Dr.Öğr.Üyesi Adem DOĞANER
Kahramanmaraş Sütçü İmam Üniversitesi
Üye

ONAY

Bu tez, İnönü Üniversitesi Lisansüstü Eğitim-Öğretim Yönetmeliği'nin ilgili maddeleri uyarınca yukarıdaki jüri üyeleri tarafından kabul edilmiş ve Enstitü Yönetim Kurulu'nun ...../...../2018 tarih ve 2018/..... sayılı Kararıyla da uygun görülmüştür.

Prof. Dr. Yusuf TÜRKÖZ
Enstitü Müdürü

İÇİNDEKİLER

TEŞEKKÜR

Akademik eğitimim ve çalışmalarımın yanı sıra günlük yaşantımda bilgi, birikim ve deneyimleri ile bana yol gösteren ve destek olan değerli danışman hocam Sayın Prof. Dr. Saim YOLOĞLU'na, eğitimim boyunca desteğini esirgemeyen ve önerileriyle bana ışık tutan değerli hocam Sayın Prof. Dr. Cemil ÇOLAK'a, aynı bölümde görev yaptığım çok kıymetli asistan arkadaşlarıma sonsuz teşekkür ve saygılarımı sunarım. Bu süreçte yardımını hiç esirgemeyen, destekleriyle beni hiçbir zaman yalnız bırakmayan aileme, eşim Can ÇİÇEK ile kızım Asya Umay ÇİÇEK'e sonsuz teşekkür ederim.

Arş. Gör. İpek BALIKÇI ÇİÇEK

ÖZET

Tip 2 Diyabet Mellitus İle İlişkili Risk Faktörlerini Saptamada Çok Değişkenli İstatistiksel Yöntemlerinin Karşılaştırılması

Amaç: Bu çalışmanın amacı, Tip 2 Diyabet Mellitus olan ve olmayan hastalara ait verileri kullanarak Yapay Sinir Ağları, Lojistik Regresyon Analizi ve Karar Ağaçları Yöntemlerine ait sınıflandırma performanslarını karşılaştırmak ve Tip 2 Diyabet Mellitus için risk faktörlerini belirlemektir.

Materyal ve Metot: Çalışmadaki veriler, İnönü Üniversitesi Tıp Fakültesi Turgut Özal Tıp Merkezi İç Hastalıkları Anabilim Dalı Diyabet ve Tiroid polikliniğine gelen hastalardan elde edilmiştir. İlgili veri seti 25 bağımsız değişken ve 1 bağımlı değişkenden oluşmaktadır. Yöntemlerin sınıflandırma performansları karşılaştırılırken performans ölçütlerinden doğruluk, duyarlılık, seçicilik, kesinlik, F-ölçümü, AUC ve sınıflama hatası kullanılmıştır.

Bulgular: Üç yöntemden en iyi sınıflandırma performansını Yapay Sinir Ağları yöntemi vermiştir. Bu yöntemin, doğruluk, duyarlılık, seçicilik, kesinlik, F-ölçümü, AUC ve sınıflama hatası değerleri sırasıyla; 98.94, 100, 97.73, 98.04, 99.01, 0.978 ve 1.06 olarak bulunmuştur. Yapay sinir ağları yöntemi için hastalığı etkileyen risk faktörlerinden cinsiyet, aile öyküsü, uzun süre ilaç kullanımı, kortizon kullanımı, eşlik eden hastalık, yüksek tansiyon, stres faktörü, kalp hastalığı, kolesterol yüksekliği, sigara kullanımı, alkol tüketimi, egzersiz durumu, karbonhidrat kullanımı, sebze kullanımı, et kullanımı, yaş, kilo, boy, başlama yaşı, günlük ekmek tüketimi, HDL, LDL, Trigliserid, Total Kolesterol, Açlık kan şekeri bağımsız değişkenlerinin ağırlıkları sırasıyla; 0.017, 0.013, 0.009, 0.008, 0.017, 0.008, 0.016, 0.024, 0.053, 0.006, 0.007, 0.023, 0.040, 0.020, 0.007, 0.046, 0.083, 0.049, 0.024, 0.066, 0.083, 0.084, 0.031, 0.020, 0.244 olarak bulunmuştur.

Sonuç: Lojistik Regresyon, Yapay Sinir Ağları ve Karar Ağaçları sınıflandırma yöntemleri uygulandığında en iyi performansı gösteren Yapay Sinir Ağları yöntemi ve bu yöneteme göre Tip 2 Diyabet Mellitusa neden olabilecek en önemli risk faktörün açlık kan şekeri olduğu elde edilmiştir.

Anahtar Kelimeler: Çok değişkenli istatistiksel yöntemler, Karar Ağaçları, Lojistik Regresyon Analizi, Tip 2 diyabet mellitus, Risk faktörleri, Yapay Sinir Ağları.

ABSTRACT

A Comparison of Multivariate Statistical Methods to Detect Risk Factors for Type 2 Diabetes Mellitus

Aim: The aim of this study was to compare the classification performance of Logistic Regression Analysis, Artificial Neural Networks and Decision Trees Methods by using data from patients with and without Type 2 Diabetes Mellitus and to determine risk factors for Type 2 Diabetes Mellitus.

Material and Method: The data in this study were obtained from patients who came to İnönü University Faculty of Medicine Turgut Özal Medical Center Internal Medicine Department Diabetes and Thyroid Polyclinic. The data set consists of 25 independent variables and 1 dependent variable. Accuracy, sensitivity, specificity, precision, F-measurement, AUC and classification error were used in the performance criteria when comparing the classification performances of the methods.

Results: Among the three methods, the best classification performance was given by the Artificial Neural Networks method. The accuracy, sensitivity, specificity, precision, F-measurement, AUC and classification error of this method were found as 98.94, 100, 97.73, 98.04, 99.01, 0.978 and 1.06, respectively. According to the results of Artificial Neural Networks method; from the risk factors affecting the disease, sex, family history, long-term drug use, cortisone use, concomitant disease, high blood pressure, stress factor, heart disease, high cholesterol, smoking, alcohol consumption, exercise status, carbohydrate use, vegetable use, meat use, age, weight , height, starting age, daily bread consumption, HDL, LDL, Triglyceride, Total Cholesterol, fasting blood sugar independent variables obtained weight values respectively; 0.017, 0.013, 0.009, 0.008, 0.017, 0.008, 0.016, 0.024, 0.053, 0.006, 0.007, 0.023, 0.040, 0.020, 0.007, 0.046, 0.083, 0.049, 0.024, 0.066, 0.083, 0.084, 0.031, 0.020, 0.244.

Conclusion: When the Artificial Neural Networks, Logistic Regression and Decision Trees classification methods are applied, the Artificial Neural Networks method has shown the best performance and according to this method, the most important risk factor that may cause diabetes is fasting blood sugar**.**

Key Words: Multivariate statistical methods, Decision trees, Logistic regression analysis, Type 2 diabetes mellitus, Risk factors, Artificial neural networks.

SİMGELER VE KISALTMALAR DİZİNİ

LRA	: Lojistik Regresyon Analizi
YSA	: Yapay Sinir Ağları
OR	: Odds Oranı
Tip 2 DM	: Tip 2 Diyabet Mellitus
kNN	: k-En Yakın Komşu algoritması
AUC	: ROC eğrisi altında kalan alan
HDL	: Yüksek yoğunluklu lipoprotein
LDL	: Düşük yoğunluklu lipoprotein
$\bar{X}$	:Aritmetik Ortalama
SD	:Standart Sapma
Min	:Minimum
Maks	:Maksimum

ŞEKİLLER DİZİNİ

Şekil No **Sayfa No**

TABLOLAR DİZİNİ

Tablo No **Sayfa No**

1. GİRİŞ

Sınıflandırma problemlerinde kullanmakta olan birçok model ve bu modellerin de çeşitli algoritmaları vardır. Modellerin algoritmalarından hangisinin daha iyi sonuçlar verdiği, hangi alanda daha başarılı olduğunu bilmek uygulamaların başarısını arttıracaktır. Bu nedenle algoritmaların karşılaştırılması ve bu karşılaştırılma sonrası değerlendirilmesi önemlidir. Fazla sayıda algoritmanın bulunması ve her birinin kendi içinde çeşitli parametreler ile çalışması, çok sayıda versiyonunun olması, farklı amaçlara yönelik olması, desteklediği veri tipinin farklı olması ve veri seti üzerindeki önişlemlerin uygulayana bağlı olmasından dolayı bu modellerden çeşitli sonuçlar elde edilmektedir (1-6).

LRA modelinde bağımsız değişkenlerin sürekli sayısal, kategorik, vb. yapıda olduğu, bağımlı değişkenin ise sadece kategorik yapıda olduğu durumda herhangi bir dağılım şartı olmadan bağımsız değişkenler ile bağımlı değişken arasında bulunan sebep-sonuç ilişkisini göstermek için tercih edilen yöntemdir. LRA, kullanılan veri setinden elde edilmiş olan olasılığı maksimum yapmak için en çok olabilirlik tahmin yöntemini kullanıp eksik parametre değerlerini tahmin etmeye çalışır (3, 7, 8).

YSA modeli çalışma yapısı olarak insan beyninde bulunan sinir ağlarına benzer bir şekilde çalışır ve karmaşık yapıdaki problemlere kolayca çözüm imkanı sunar. Modelde her türde değişken kullanılır ve değişkenlerin birbirleri ile olan ilişkisini gösterir. YSA modeli sınıflandırma problemlerinde yaygın olarak tercih edilmektedir. Sınıflandırma yaparken algoritma olarak geriye yayılma yöntemini kullanıp ağ hatasını en aza indirerek hesaplama yapar (3, 9).

Karar ağacı modelinde verilerin sınıflandırılması iki adımda yapılmaktadır. İlk adım; bilinmekte olan bir eğitim verisi ile modeli meydana getirmek için sınıflandırma algoritması yardımıyla çözümlemeler yapılır ve öğrenme gerçekleşir. Öğrenilmiş olan model, karar ağacı veya sınıflandırma kuralları ile gösterilir. Diğer adımda eğitim verisinin karar ağacının veya sınıflama kurallarının doğruluğunu belirleyebilmek için test edildikten sonra kullanıldığı bir sınıflamadır (10).

Literatür taraması sonucunda çalışmada uygulanacak olan üç sınıflandırma modelinden çoğunlukla ikişerli karşılaştırılma yapılmıştır. Yapılan taramanın sonucunda güncel çalışmalardan bazılarına yer verilmiştir.

Güneri ve ark. (3) YSA ve LRA modelleri kullanılarak yaptığı çalışmada sınıflandırma oranları iki model için %95.17 olarak elde edilmiştir. İki modelin eşit sonuç vermesi YSA modelinin sınıflandırma problemlerinde kullanılabileceğinin göstergesi olmuştur.

Karakış'ın (11) çalışmasında meme kanseri olan hastaların koltuk altındaki lenf nod durumunun belirlenebilmesi için YSA modeli kullanılmıştır. Çalışmada Ankara da bulunan iki araştırma hastanesine başvuran ve meme kanseri olan 270 kişiden toplanan veriler kullanılmıştır. YSA modelinde korelasyon katsayıları ve regresyon sonuçları değerlendirilmiştir. YSA modeli sonuçlarının karşılaştırılmasında LRA modeli kullanılmıştır. LRA da verilerin anlamlılığına bakılmış ve anlamlı olanlar YSA modelinde tekrar eğitilerek test edilmiştir. Sonuç olarak YSA modeli değerleri daha başarılı bulunmuştur.

Bu çalışmada amaç bağımlı değişkenin yapı olarak kategorik olduğu verilerin analizi yapılırken bağımsız değişkenlerin bu değişkene olan etkilerini sorgulamak ve hata düzeyini en aza indirerek sınıflandırılmada kullanılmakta olan Lojistik Regresyon Analizi (LRA), Yapay Sinir Ağları (YSA) ve Karar Ağaçları modellerinin karşılaştırılmasını yapmaktır ve bu modeller kullanılarak bağımlı değişkene etki eden risk faktörleri incelenmektir.

2. GENEL BİLGİLER

Tahmin etmede kullanılan modellerin amacı, araştırma verileri yardımıyla bir model oluşturmak ve oluşturulmuş bu model ile veri kümelerinde bilinmeyen sonuçların değerlerini tahmin edebilmektir. Eğer tahmin edilmesi gereken değişken sürekli yapıda bir değişken olduğunda problem regresyon problemi olarak ele alınırken kategorik yapıda bir değişken olduğunda sınıflandırma problemi olarak ele alınmaktadır (2). Bu çalışma da sınıflandırma modellerinden Lojistik Regresyon, Yapay Sinir Ağları ve Karar Ağaçları kullanılacaktır.

2.1. Lojistik Regresyon Analizi (LRA)

Lojistik regresyon; cevap değişkeni (Y) yani bağımlı değişken türünün iki kategorili (binary) ve çoklu (multinominal) kategorili olduğu durumlarda açıklayıcı değişken (X_i,i=1,2,…,k) olarak bilinen bağımsız değişkenlerle sebep sonuç ilişkisi belirlemede kullanılan bir yöntemdir. Lojistik regresyon modellerinin tıp alanındaki uygulamalarında bağımsız değişkenler, risk değişkenleri ya da bir hastalığın ortaya çıkıp çıkmamasını belirleyen değişkenlerdir. Örneğin sigara içip içmeme akciğer kanserine yakalanıp yakalanmamayı belirleyen bir değişkendir. O halde buradaki risk değişkeni sigara değişkenidir. Lojistik regresyon, bağımlı değişkenin bağımsız değişkenlere göre beklenen değerinin ifade edildiği sınıflandırma ve atama işlemi yapan regresyon yöntemlerinden biridir.

Gözlemlerin gruplara ayrılmasında kümeleme analizi, lojistik ve ayırma analizi yöntemleri kullanılır. Kümeleme analizine göre değişkenlerin atamasının yapılacağı grup sayısı tam olarak bilinemez ve değişkenler benzerlik-uzaklık ölçülerine bakılarak gruplandırılır. Lojistik regresyon ile ayırma analizinde ise grup sayısı önceden belirlidir ve ayırsama modeli bulunur.

Basit ve çoklu regresyon analizlerinin uygulanabileceği verilerde bağımlı ve bağımsız değişkenlerin bazı varsayımları yerine getirmesi gerekir. Bu varsayımlardan bazıları bağımlı değişkenin normal dağılması, bağımsız değişkenin de normal dağılması ve hataların terimlerinin varyansı sabit ve normal dağılıma uygun olmalıdır. Basit ve çoklu regresyon analizleri bu ve benzeri varsayımların sağlanmadığı veri setlerinde uygulanamaz. Lojistik regresyon analizi ise süreklilik varsayımı ve normal dağılım varsayımı ön koşulu olmayan bir regresyon yöntemidir. Bağımlı değişken üzerinde etkisi olan bağımsız değişkenlerin etkilerini olasılık olarak belirlemeyi sağlayan bir yöntemdir (12-14).

Sağlık alanında çalışan araştırıcılar çalıştıkları konuda birden çok etkenin olduğu durumda etkenlerin bağımlı değişken üzerine etkisini tek tek öğrenmenin yanında, bunların birlikte bağımlı değişken üzerinde etkisini de bilmek ve incelemek istemektedirler. Birlikte etkinin araştırılmasında kullanılan birçok istatistiksel yöntem vardır. Bağımsız değişkenlerin kesikli, bağımlı değişkenin sürekli olması durumunda varyans analizi, bağımlı ve bağımsız değişkenlerin kesikli olması durumda "log-linear model"ler, hem bağımlı değişkenin hem de bağımsız değişkenlerin sürekli yapıda olduğu durumunda ise regresyon analizinin kullanılması örnek olarak verilebilir. Tıp alanında yapılan çalışmalarda çoğunlukla bağımsız değişkenlerin ve bağımlı değişkenlerin türü ve yapıları yukarıdaki örnekte bahsedilenlere benzemeyebilir, değişkenler sürekli ve kesikli olabilir. Araştırıcı için önemli olan bir diğer konu da hastalığın bir etkenle ilişkisini risk yönünden incelemektir. Bu tür araştırmalarda ağırlıklı olarak Lojistik regresyon analizi kullanılmaktadır (15, 16). Son zamanlarda çok sık kullanılan Lojistik regresyon analizi, gözlemlerin gruplara atanmasında kullanılan yöntemlerden birisidir. Grup sayısı bilinen lojistik regresyon analizinde veriler kullanılarak ayırsama modeli elde edilir ve elde edilen bu model veriye eklenen yeni gözlemlerin gruplara atanmasını sağlayabilmektedir (17, 18).

Doğrusal regresyonda bağımlı değişkenin değerinin kestirilmesi ile ilgilenilirken lojistik regresyonda bağımlı değişkenin değerlerinden birinin gerçekleşme olasılığı kestirilmeye çalışılır. Olasılık değerinin kestirimi yapılırken kullanılan model aşağıdaki gibidir.

$$\pi(x) = \frac{e^{\beta_0+\beta_1 x}}{1 + e^{\beta_0+\beta_1 x}}$$

Lojistik regresyon modelinde regresyon katsayıları tahmin edilirken genellikle kullanılan yöntem en çok olabilirlik yöntemidir. Bu yöntem, gözlenen veriyi elde edilme olasılığını en büyük yapan eksik parametrelerin değerlerini üretmede kullanılır. Bu yöntem uygulanırken en çok olabilirlik fonksiyonunu olarak bilinen fonksiyonun oluşturulması gerekmektedir. Bu fonksiyon, gözlenen verideki olasılığı eksik olan parametrelerin bir fonksiyonu olarak gösterir. Bu yöntemde bir olayın gerçekleşme olasılığı maksimum yapılmaya çalışılır. Lojistik regresyon analizinde bağımsız değişkenlerle bağımlı değişken arasındaki ilişkinin önemli olup olmadığı incelenir. Bunun için değişken modelde var iken ve değişken modelde yok iken elde edilen tahmin değerleri ile gözlenen değerler karşılaştırılır. Değişken modelde var olduğunda daha iyi, daha doğru tahminler elde edilirse değişkenin

modelde önemli olduğu sonucuna varılır. Modeldeki değişkenlerin önemli olup olmadığı Olabilirlik Oranı, Skor ya da Wald testlerinden biri ile bakılabilir (12).

2.1.1. Olabilirlik Oranı Testi

Kurulan modelde bağımsız değişkenin önemliliği için karar verilirken; denklemde bağımsız değişkenin yer aldığı ve almadığı durumlar göz önüne alınarak bu durumlardaki değerler olabilirlik oranı olarak bilinen G istatistiği ile karşılaştırılır.

$$G = -2\ln\left(\frac{L(değişken\ modelde\ olmadığında)}{L(değişken\ modelde\ olduğunda)}\right)$$

$$G = -2\ln L((değişken\ modelde\ olmadığında) - \ln(değişken\ modelde\ olduğunda))$$

G istatistiği asimtotik olarak ki-kare dağılımı gösterir. Bu test olabilirlik oranı testi olarak adlandırılır (12).

2.1.2. Skor Testi

Skor testinde en çok olabilirlik tahmininin hesaplanmasına ihtiyaç yoktur. Bu nedenle, hesaplama işlemlerini çok fazla kısalmakta bu da bu testin en büyük avantajı olmaktadır. Skor testi, çok değişkenli olup matris hesaplamaları gerektirir. Aşağıdaki eşitlik ile verilir.

$$ST = \frac{\sum_{i=1}^{n} x_i(y_i - \bar{y})}{\sqrt{y_i(1-\bar{y})\sum_{i=1}^{n}(x_i - \bar{x})^2}}$$

Skor test istatistiği normal dağılıma uymaktadır (12).

2.1.3. Wald Testi

Wald testinde olabilirlik oran testindeki gibi en çok olabilirlik tahmininden yararlanılır. Wald testi, eğim parametresi olan β_1'nin en çok olabilirlik tahmini olan $\widehat{\beta_J}$' nin standart hatasına $S(\widehat{\beta_J})$'ye bölünmesi ile bulunur. Aşağıdaki eşitlik ile verilir.

$$W = \frac{\widehat{\beta_J}}{S(\widehat{\beta_J})}$$

Wald test istatistiği normal dağılıma uymaktadır.

Modeldeki değişkenlerin önemlilikleri test edildikten sonra sonuçların yorumlanması gerekmektedir. Bunun için risk kavramı ve risk tahminine ilişkin bazı ölçülerin bilinmesi gerekir. Bu ölçüler göreli risk ölçüleri ve genel risk ölçüleridir (12).

2.1.4. Genel Risk Ölçüleri

Bir hastalığın görülme olasılığı evrendeki riskini de tanımlar. Genel risk ölçüsü olan insidans ve prevalans riski tanımlamak için kullanılır. İnsidans hızı; sağlıklı kişilerden alınan grubun bir süre geçtikten sonra yüzde kaçının hastalık geçirdiğine bakılır. Yani; insidans ortaya çıkan yeni olgularla ilgilidir. Bu sebepten dolayı atak hızı da denir.

İnsidans Hızı=Yeni Olgu Sayısı / Risk Altındaki Nüfus

Prevalans hızı ise hem eski hem de yeni olgularla ilgilenir.

Prevalans Hızı=(Eski+Yeni Olgu Sayısı) / Risk Altındaki Nüfus

2.1.5. Göreli Risk Ölçüleri

Bir etkenle hastalığın ilişkisini gösterir. Hastalığın ortaya çıkmasındaki etkenleri (risk faktörlerini) göz önünde bulunduran risk ölçüleri Göreli Risk (RelativeRisk-RR) ve Odds Oranıdır (Odds Ratio-OR) (12).

2.1.5.1. Göreli Risk Kestirimi

Benzer özellikleri bulunan sağlıklı bireylerden oluşmuş iki grup rastgele seçilir. Yani; etkenin bulunduğu gruptan n_1, etkenin bulunmadığı gruptan n_2 kişi rastgele seçilir ve bu iki grup belirli bir zaman takip edilir. İzlenen süre sonunda etkenin etkisinde bulunan ve bulunmayan gruplardaki insidans hızları hesaplanır. Göreli risk; etkenin bulunduğu gruptaki riskin, etkenin bulunmadığı gruptaki riske bölünmesi ile bulunur (12).

2.1.5.2. Odds Oranı (Göreceli olasılıklar oranı) Kestirimi

Hastalığın bulunduğu gruptan n_1, hastalığın bulunmadığı gruptan n_2 kişilik örneklem alınır, kişilerde hastalığın üzerine bir etkisinin bulunduğu düşünülen etkenin olup olmadığına bakılır. Odds oranı için tablo gösterimi aşağıdaki gibidir.

Tablo 2.1. Odds Oranı için Tablo

Etken	Hastalık		Toplam
	Var	Yok	
Var	a	b	a+b
Yok	c	d	c+d
Toplam	a+c=n_1	b+d=n_2	n

Tablo 2.1'den

Hasta grupta odds: a / c

Sağlıklı grupta odds: b / d

olarak tanımlanır.

İki odds'un oranına Odds Oranı (OR)

$$OR = \frac{\text{a / c}}{\text{b / d}} = \frac{a \times d}{b \times c}$$

OR= 1 olduğu durumda etkenin hastalığın riskini artırıcı veya azaltıcı bir etkisi yok denilebilir.

OR < 1 olduğu durumda etkenin hastalığın riskini azaltıcı olduğu söylenirken,

OR > 1 ise olduğu durumda etkenin hastalığın riskini artırıcı olduğu söylenir (12).

2.2. Lojistik Sınıflandırma ve Lojistik Regresyon Modeli

İki kategorili olan lojistik regresyonda bağımlı değişkenin 0 ve 1 kodlanan değerlerine G_1 ve G_2 grupları karşılık gelsin. Bu gruplar $x_1, x_2, \ldots, x_p$ bağımsız değişkenlerine bağlı olarak sınıflandırılmak istensin. G_1 grubundaki birey sayısı n_1 ve G_2 grubundaki birey sayısı n_2 olmak üzere; $N = n_1 + n_2$ gözlem için sınıflandırma kuralını oluşturmak olasılık fonksiyonu olan $f_s(x_1, x_2, \ldots, x_p)$ şeklindeki fonksiyonun fonksiyonel yapısından dolayı ortaya çıkan şartlara bağlıdır. Bu olasılık fonksiyonu için söz konusu olan üç farklı varsayım vardır (4, 5, 14).

i. Lojistik sınıflandırma fonksiyonu
ii. Çok değişkenli normal dağılım fonksiyonu

iii. Dağılıma bağlı olmayan çekirdek sınıflandırma fonksiyonu

Söz konusu olan lojistik sınıflandırma fonksiyonu ise $x_0 = 1$ iken $f_s(x_1, x_2, \dots, x_p)$, G_s $(s = 1,2)$ olasılık yoğunluk fonksiyonudur. Lojistik varsayım, $\beta' = (\beta_0, \beta_1, \dots, \beta_p)$ için,

$\frac{f_1(x)}{f_2(x)} = \exp(\beta' X')$ ya da da $\ln\left(\frac{f_1(x)}{f_2(x)}\right) = \beta' X'$

şeklinde tanımlanır. Bu eşitlik log-olabilirlik oranı olarak tanımlanır ve *x*' ler doğrusaldır. Bilinmeyen β parametreleri lojistik varsayım içermektedir. Sonsal olasılık X şartı altında gözlemlerin her biri için gruplardan herhangi birine atanma olasılığı şeklinde kabul edilir. Sonsal olasılıkları hesaplamakta β tahminleri gereklidir. Örneklemin olabilirlik fonksiyonu lojistik varsayıma göre belirlenmelidir. Karışık örneklemede gözlemler rasgele seçilir hangi gruptan geldiği bilinmez yani gözlemler (X,G) bileşik dağılımdan örneklenir. Burada; G değişkeni grup üyeliğini gösterir (19).

Koşullu örnekleme durumunda ise varsayılan x koşulu altında G' nin nasıl dağıldığı incelenmektedir. Bu örnekleme türü ile ilişkili olan olabilirlik fonksiyonu biyolojik deneylerin analizinde çok sık kullanılır ve diğer örnekleme türleri ile ilgili olabilirlik fonksiyonlarına da temel olarak gösterilir (16).

Anahtar değer; regresyon problemlerinde bağımlı değişkenin ortalama değeri verilen bağımsız değişken değerine bağlı olarak bulmaktır. Bu değere koşullu ortalama adı verilir ve $E(Y\backslash \mathrm{x})$ ile gösterilir ve *x* verildiğinde, y' nin beklenen değerini ifade etmektedir. Ayrıca bu ifadede de y; bağımlı değişkeni *x* ise bağımsız değişkeni gösterir. Koşullu ortalamanın doğrusal regresyon analizinde, *x*' e ait doğrusal bir denklem ile ifade edildiği varsayılır.

$$E(Y\backslash \mathrm{x}) = \beta_0 + \beta_1 x$$

Yukarıdaki ifadede, *x*'in alacağı değerler -∞ ve +∞ arasında değiştiği için $E(Y\backslash \mathrm{x})$' in her değeri alabileceği görülmektedir. Bağımlı değişken iki kategorili olduğunda koşullu ortalamanın değeri sıfır ile bir arasında değer almak zorundadır (19). $0 \leq E(Y\backslash \mathrm{x}) \leq 1$. $E(Y\backslash \mathrm{x})$' deki değişiklik, *x*' deki her birim değiştikçe, koşullu ortalama değeri sıfıra veya bire yaklaştıkça azalır. İki düzeyden oluşan bir bağımlı değişkenin analizi yapılırken kullanılması önerilen bazı dağılım fonksiyonları bulunmaktadır. Lojistik dağılımda $\pi(x) = E(Y\backslash \mathrm{x})$ ifadesi

x biliniyorken Y' nin koşullu ortalaması anlamına gelir. Lojistik regresyonda kullanılacak olan modelinin açık şekli aşağıda gösterilmiştir (8, 20, 21).

$$\pi(x) = \frac{e^{\beta_0+\beta_1 x}}{1 + e^{\beta_0+\beta_1 x}}$$

$\pi(x)$'in transformasyonu yukarıda bahsedildiği gibi bir lojit transformasyonudur. $\pi(x)$ cinsinden bu transformasyon tanımlanacak olursa:

$$g(x) = \left[\ln \frac{\pi(x)}{1-\pi(x)}\right] = \beta_0 + \beta_1 x$$

Yukarıda gösterilen lojit $g(x)$ parametreleri açısından doğrusal olup x' in alacağı değerlere göre $-\infty$ ve $+\infty$ aralığında değişmektedir (8, 20, 21).

2.2.1. LRA Modelinin Elde Edilmesi

Doğrusal regresyon modelinde olduğu gibi LRA modelini oluşturmak için de, maksimum olabilirlik tahmin metodu kullanılmaktadır. n tane bağımsız gözlem ikilisi (x_i, y_i) olsun. Bu ikili de y_i iki kategorili bağımlı değişken olmak üzere ve x_i' de i' inci deneğin bağımsız değişkeninin değerini gösterirken; cevap değişkeni için ise 0 ve 1 kodlarının sırasıyla belirli bir olayın varlık veya yokluk durumunu gösterdiğini varsayalım. LRA modelini tahmin etmek için;

$$E(Y\backslash x) = \beta_0 + \beta_1 x$$

yukarıdaki eşitlikte bilinmeyen β_0 ve β_1 parametrelerini tahmin etmek gerekir. Eğer cevap değişkeni olan Y, 0 ve 1 olacak şekilde kodlandıysa $\pi(x)$ ifadesinin anlamı; x durumu verildiğinde Y' nin l'e eşit olma durumundaki koşullu olasılığıdır. Herhangi bir x durumunda Y olasılığının 0'a eşit olma koşullu olasılığını [$\pi(x) = P(Y = 1\backslash x)$] . [1- $\pi(x)$] değeri göstermektedir.

[1- $\pi(x)$ = P(Y = 0\ x)] . (x_i, y_i) çiftinde $y_i = 1$ olduğu durumda olabilirlik fonksiyonunun değerine $\pi(x_i)$ kadar bir katkısı varken, $y_i = 0$ olduğunda olabilirlik fonksiyonunun değerine 1-$\pi(x_i)$ kadar katkısı olur. (x_i, y_i) den oluşan ikilinin olabilirlik fonksiyonuna var olan katkısı;

$$\xi(x_i) = \pi(x_i)^{y_i}[1 - \pi(x_i)]^{1-y_i}$$

ifadesi yardımıyla tanımlanır (16, 19).

Olabilirlik fonksiyonunu; gözlemlerin birbirinden bağımsız olduğu varsayıldığı için,

$$E(Y\backslash x) = \beta_0 + \beta_1 x$$

eşitliğinde bulunan terimlerin çarpılmasıyla bulunur.

$$\int(\beta) = \prod_{i=1}^{n} \xi(x_i)$$

β tahmininin $g(x) = \left[\ln \frac{\pi(x)}{1-\pi(x)}\right]$ eşitliğini maksimum yaptığı en çok olabilirliğin temel ilkesinde vurgulanır, ek olarak log olabilirlik fonksiyonu ise:

$$L(\beta) = \ln[1(\beta)] = \sum_{i=1}^{n} \{y_i \ln[\pi(x_i)] + (1 - y_i) \ln[1 - \pi(x_i)]\}$$

$L(\beta)$'nın β_0 ve β_1'e göre türevini alıp bulunan ifadedeler 0' a eşitlenirse $L(\beta)$' yı maksimum yapacak β değerini bulunur. Sonunda bulanan en çok olabilirlik eşitlikleri aşağıdaki gibidir (16, 19).

$$\sum_{i=1}^{n} [y_i - \pi(x_i)] = 0$$

$$\sum_{i=1}^{n} x_i [y_i - \pi(x_i)] = 0$$

2.2.2. LRA Modelindeki Katsayı Yorumu

Yapılması gereken ilk adım bağımsız değişkenler ile bağımlı değişkenin hangi fonksiyonunun doğrusal fonksiyon oluşturduğunun belirlenmesidir. Bu oluşan doğrusal fonksiyona link fonksiyonu denir. Doğrusal regresyon modelinde oluşan link fonksiyonu birim matris olup, bunun nedeni doğrusal regresyon modelinde bağımlı değişkenin parametreleri ile doğrusallık göstermesidir. LRA modelinde oluşan link fonksiyonu ise lojit transformasyonudur (16, 18).

$$g(x) = \left[\ln \frac{\pi(x)}{1 - \pi(x)}\right] = \beta_0 + \beta_1 x$$

2.2.3. Modelde İkili Bağımsız Değişkenin Olduğu Durum

Lojistik regresyon modelinde katsayıların yorumuna bağımsız değişkenin ikili olduğu durum ile başlanacaktır. Bağımsız değişkenin ikili olduğu durum en basit durum olup diğer durumlar için bir temel olacaktır. *x*'in 0 ve 1 olarak kodlandığını varsayalım. Bu model için π (*x*)'in ve 1 - π (*x*) 'in iki değeri vardır. π *(x)* için bu iki değer π (0) ve π (1) iken, 1 - π (*x*) için bu iki değer 1 - π (0) ve 1 - ve π (1)'dir (16, 18).

2.2.4. Modelde Sürekli Bir Bağımsız Değişkenin Olduğu Durum

Lojistik regresyon modelinde bağımsız değişkenin sürekli bir değişken olduğunda katsayının yorumlanması için geliştirilecek yöntemde değişken ile lojitin doğrusal olduğu varsayılacaktır. Sürekli değişken ile lojitin (*x*) doğrusal olduğu varsayıldığında lojit için eşitlik $g(x) = \beta_0 + \beta_1$'dir. Eğim katsayısı olan (β_1); *x*'deki "1" birimlik artış ile log odds değerinde meydana gelecek değişimi ifade eder. Log odds değeri *x*'deki "c" birimlik değişim için lojit farktan elde edilir. $g(x + c) - g(x) = c\beta_1$. Lojit farkın üssü alınarak karşılık gelen odds oranı elde edilir. . $\psi(c) = \psi(x + c, x) = \exp(c\beta_1)$ $\psi(c)$'nin tahmininde güven aralığının uç noktaları

$$\exp\left[c\widehat{\beta_1} \pm z_{1-\alpha} c\hat{S}E\left(\widehat{\beta_1}\right)\right]$$

olarak verilmiştir (21).

2.3. Yapay Sinir Ağları (YSA)

İnsanlar bilgisayarlar üzerinde beynin çalışma şeklini taklit etmek istemişler ve YSA ortaya çıkmıştır. İnsanlar gibi düşünebilen ve öğrenebilen sistemler oluşturma fikrinin olması, yapılacak çalışmaları beynin çalışma yapısının nasıl olduğu fikrine yöneltmiştir. Bu çalışmalar sırasında bütün nöronların ilişki halinde olduğu ve aldığı girdileri çıktı haline getirdiği görülmüştür. Bu alan günümüzde YSA ile ifade edilmektedir ve belirli kurallar ile nöronların bir araya gelip bir fonksiyonun gerçekleştirmesi sırasında matematiksel, yapısal ve felsefi sorunlara yanıt arayan bir bilim dalıdır (22).

YSA, beynin herhangi bir görevini ya da işlevini yerine getirme metodunu model haline getiren bir makinedir. Bilgisayar yazılımları veya elektronik olan sistemler ağın yapısını tanımlamaktadır. İşlem birimi veya sinir hücresi olarak adlandırılan bu model verilen hücreler arasındaki bağlantıyı kullanır. Bu işlemleri yerine getirirken öğrenme olarak adlandırılan süreç ile performansını artırmaktadır. Bu tanımdan yola çıkarak yapay sinir ağı kavramı bilgiyi saklayan ve kullanmak için hazır halde bulunduran basit işlem birimlerinden meydana gelen paralel, aşırı yoğun ve dağınık bir düzende çalışma sistemine sahip olan işlemci olarak ifade edilebilir. Yapay sinir ağlarının sinir hücrelerinin arasında bulunan bağ ile bilgiyi depolaması ve öğrenme yolu ile bilgiyi elde etmesi yönüyle insan beyni ile benzerdir (23).

2.3.1. YSA Nedir?

YSA, bir bilgi işleme sistemi olup biyolojik sinir ağlarının özelliklerine benzer özelliklere sahiptir. Yapay sinir ağları biyolojik nöron yapısının matematiksel olarak modelin aşağıdaki kuralların genel duruma getirilmesi sonucunda oluşturulmuştur:

- Bilgi işleme, nöron olarak adlandırılan elemanlarda gerçekleşir;
- Sinyaller, nöronların birbiri ile iletişimini sağlayan bağlantılarla iletilir;
- Bağlantıların her birinde ağırlık değeri vardır ve bağlantıların sahip olduğu bu değer, beyindeki nöronlar gibi sinyal geçişini üretir;
- Sinir ağının içerisinde bulunan nöronların her birine genelde doğrusal olmayan aynı aktivasyon fonksiyonu uygulanır ve bu aktivasyon fonksiyonundaki çıkış değeri ile nöronun çıkış sinyali değeri hesaplanır;

YSA;

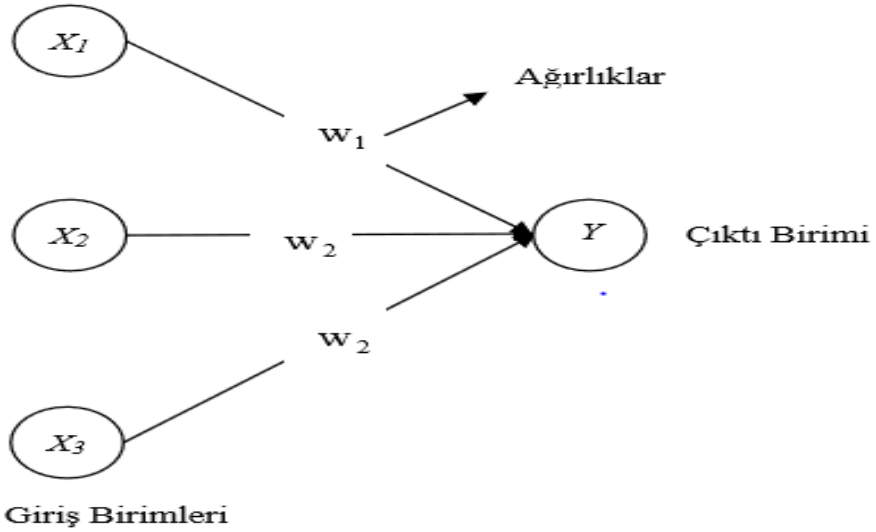
- Nöronların birbiri arasında bulunan bağıntının şekli ile,
- Öğrenme algoritması ya da eğitim kuralı olarak adlandırılan bağlantılardaki ağırlıkların hesaplanması
- Aktivasyon fonksiyonu ile açıklanabilir.

YSA; birim, hücre veya düğüm ve nöron olarak adlandırılan işlem birimlerinden oluşmaktadır. Nöronların her biri, bir diğer nörona haberleşme bağlantıları ile bağlanır. Bu haberleşme bağlantıları belirli bir ağırlık değerine sahiptir. Ağırlıklar ise YSA' nın problemi çözebilmesi için gerekli olan bilgiyi hazır hale getirmektedir. YSA birçok problemin çözümü için kullanılabilmektedir. Örneğin, numunelerin ve bilgilerin saklanmasında ve daha sonra numunelerin ve bilgilerin tekrar tanınmasında, numunelerin sınıflandırmasında, giriş numunelerinin çıkış numuneleri olarak dönüştürülmesinde, benzeyen örneklerin gruplandırılmasında ve daha birçok alanda YSA geniş biçimde kullanılabilmektedir (24).

Her bir nöronun bir iç durumu olup bu iç durum aktivasyon veya aktivasyon düzeyi olarak isimlendirilir. Bu aktivasyon düzeyi, alınmış olan giriş değerlerinin bir fonksiyonu olur. Herhangi bir nöron, diğer nöronlara kendi aktivasyon düzeyini, genelde sinyal şeklinde gönderir. Ancak aynı anda birden fazla nörona bu sinyal gönderilebilir (25).

Örnek verecek olursak şekil-1'de gösterilmiş olan bir Y nöronunu ele alalım. Y nöronu giriş sinyallerini $X_{1,}X_2$ ve X_3 nöronlarından alır. Bu nöronların çıkış sinyalleri yani aktivasyonları, sırasıyla $x_{1,}x_2$ ve x_3 ile gösterilir. Bağlantıların üzerinde bulunan ağırlıklar bu nöronlardan Y nöronuna doğru olacak şekilde sırasıyla w_1, w_2 ve w_3'tür. Ağ girişi olarak ifade edilen y _ in değeri $X_{1,}X_2$ ve X_3 nöronlarından Y nöronuna doğru giden ağırlıklı sinyaller toplanarak bulunur. Ağ girişi olan y _ in değerini aşağıda gösterilen eşitlikteki gibi hesaplanır (24).

y _ in$= w_1x_1 + w_2x_2 + w_3x_3$



Şekil 2.1. Basit olan bir yapay nöron gösterimi

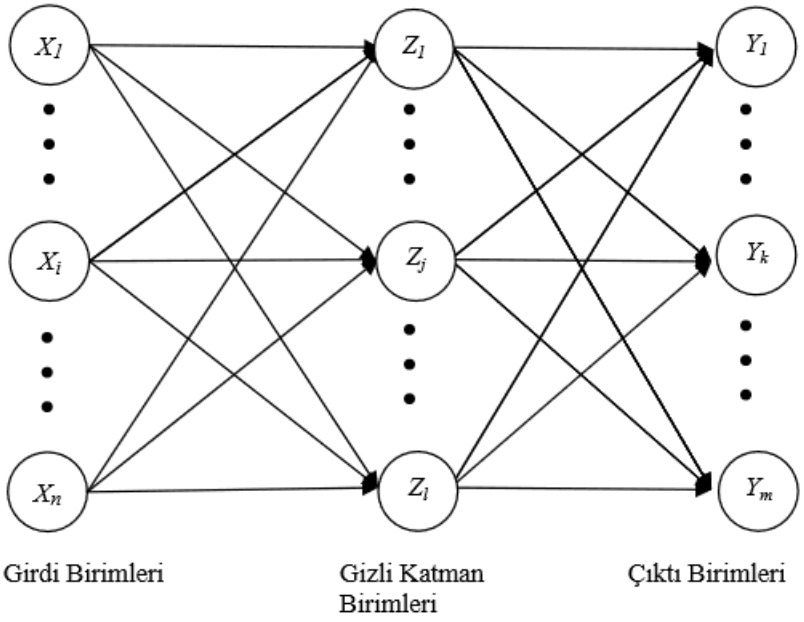
Aktivasyon fonksiyonu y olan Y nöronu, ağda bulunan giriş değerlerinin fonksiyonu ile ifade edilmektedir.

f(y _ in)=y

Buradaki aktivasyon fonksiyonu, bir sigmoid fonksiyonu olabilir ve bu fonksiyonun formülize hali aşağıdaki gibi ifade edilir.

$$f(x) = \frac{1}{1 + \exp(-x)}$$

Y nöronu v_1 ve v_2 ağırlıkları ile Z_1ve Z_2 nöronlarına bağlı olduğunu düşünülsün. Bu durum aşağıdaki gibi Şekil-2.2 'de gösterilsin. Diğer birimlere Y nöronu y sinyalini gönderir. Genelde Z_1 ve Z_2 nöronlarının aldığı sinyaller birbirinden farklı olur; bunun nedeni ise her bir sinyalin aktarılmış olduğu bağlantıda var olan v_1 ve v_2 ağırlıkları ile orantılı olacaktır. Z_1 'in aktivasyon değeri z_1 ve Z_2 'nin aktivasyon değeri z_2 olmak üzere bunlar yalnızca bir tek nörona bağlantılı değildir. Birbirinden farklı olacak şekilde birden çok nörondan gelecek olan sinyallere bağlıdır (24).

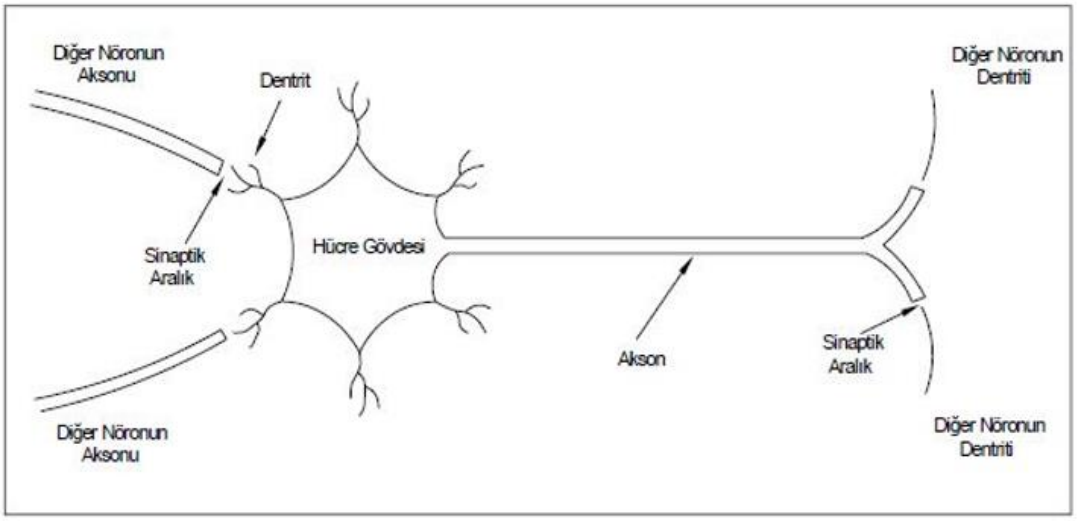


Şekil 2.2. Basit yapıda olan bir YSA

2.3.2. Sinir Ağlarının Biyolojik Yapısı

Şekil 2.3. Sinir Sisteminin Biyolojik Yapıda Gösterimi

Sinir sisteminin biyolojik yapısı beyin merkezli olup 3 katmanlı bir yapıdır. Bunlar verinin alınması, yorumlanması ve karar üretilmesi işlevleridir. Gelen uyarılar alıcı sinirleri yardımıyla elektriksel sinyallerine dönüştürülüp beyine gönderilir. Beyinde meydana gelen çıktılar, tepki sinirleri yardımıyla da belirli tepkiler haline getirilmektedir.



Şekil 2.4. Biyolojik Yapıdaki Sinir Hücresi

Sinir ağının yapısına göre bilgiler alıcı sinirlerin ve tepki sinirlerinin arasında ileri ve geri beslemeli olarak değerlendirilir. Bu değerlendirmenin sonucu olarak meydana gelen tepkiler kapalı olan bir döngü süreci ile benzetilmektedir. Temel işlem elemanı sinir hücreleridir. Şekil 2.4'de görüldüğü gibi sinir hücreleri 3 temel yapıdan oluşmaktadır. Bunlar; hücrenin gövdesi, gövdeye gelmekte olan alıcı dentritler ve gövdeden çıkmakta olan sinyal ileten aksonlardır. Dentritler yardımıyla bilgilerin hücre gövdesine iletilmesi sağlanır. Hücrelerde meydana gelen çıktılar bir hücreden diğer bir hücreye akson yardımıyla aktarılır. Bu aktarım meydana geldikten sonra aksonlardan yollara ayrılarak diğer bir hücrede yer alan dentritleri oluşturur. Aksonlarla dentritlerin bağlandığı bu noktaya sinaps denir. (9, 23).

YSA gelişimi insan beyninin çalışma prensibi göz önünde bulundurularak geliştirildiği için ikisi arasında benzerlikler bulunmaktadır (18). Bahsedilen benzerlikler aşağıdaki Tablo 2.2'de gösterilmiş olup Tablo2.3 ile de YSA modelinde kullanılan istatistiksel terimler ile modele ait olan terimlerinin terminolojik açıdan ilişkisi gösterilmiştir (18, 24).

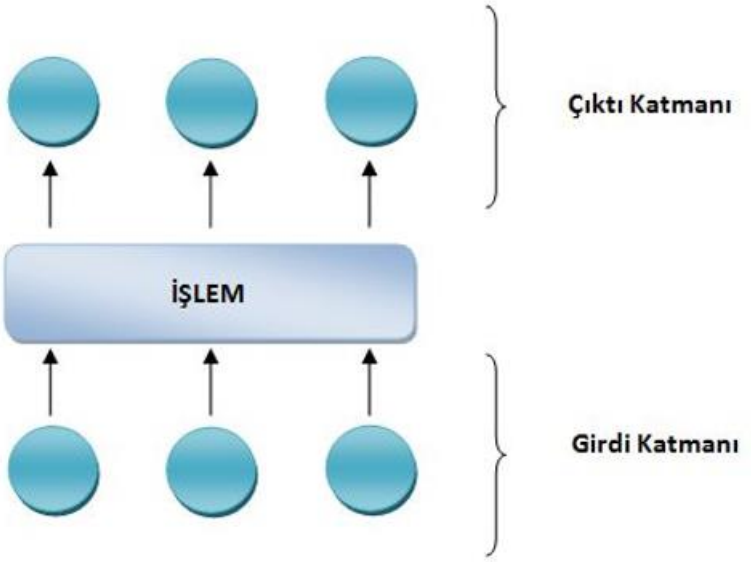
Tablo 2.2. Sinir Hücresinin Biyolojik Yapısı (18, 24)

Sinir Sistemi	Yapay Sinir Ağı
Nöron	İşlemci Eleman
Dentrit	Girdiler
Hücre Gövdesi	Transfer Fonksiyonu
Akson	Yapay Nöron Çıkışı
Sinaps	Ağırlıklar

Tablo 2.3. YSA'nın İstatistiksel Yöntemler İle Benzeşimi (18, 24)

İstatistik	Yapay Sinir Ağı
Model	Ağ
Tahmin	Öğrenme
Regresyon	Danışmalı Öğrenme
İnterpolasyon	Genelleştirme
Gözlem	Öğrenme Algoritması
Parametre	Ağ Parametreleri
Bağımsız Değişken	Giriş Verileri
Bağımlı Değişken	Çıkış Verileri
Sınır Regresyonu	Ağırlık Budama İşlemi

2.3.3. YSA' da Genel Yapı



Şekil 2.5. YSA'nın Genel Yapısı

- **Girdi Katmanı:** Girdi elemanının olduğu bu bölüme girdi katmanı denir. Girdi katmanındaki veriler girdileri ile aynı değerde çıktı üretirler yani herhangi bir işleme tabi tutulmazlar.
- **Çıktı Katmanı:** En az bir çıktıdan meydana gelen bu bölüme çıktı katmanı denir ve çıktı ise ağ yapısındaki fonksiyona bağlıdır. Bu birimlerde işlem meydana gelir ve buradaki birimler kendi çıktılarını meydana getirirler.
- **İşlem Katmanı:** Genelde "Kara Kutu" şeklinde bilinmektedir. Girdi elemanlarına belli başlı işlemlerin yapıldığı yerdir. Seçilmiş olan ağ yapısına bağlı olarak bu katmanının fonksiyonu ile yapısı da değişebilmektedir. Sadece bir katmandan meydana gelebileceği gibi birçok katmandan da meydana gelebilir (23).

2.3.4. YSA'yı Oluşturan Başlıca Elemanlar

YSA aşağıda belirtilmiş olan bilgiler üzerine inşa edilmiştir: (18, 26).

- Bilgi nöronlarda gerçekleştirilir
- Nöronlar arasındaki bağlantılardan işaretler geçer
- Her bir bağlantının bir ağırlığı olup bağlantılar birçok işareti taşıyabilir
- Her bir nörona sahip olduğu giriş değerinin çıkış işleminden sonraki işaretini belirlemesi için bir aktivasyon fonksiyonu uygulanır.

YSA insan beyninin sahip olduğu hatırlama, öğrenme ve bilgiyi genelleme yeteneğine sahiptir.

Öğrenme üç farklı şekilde olmaktadır;

- Aksonlar üreterek,
- Aksonları uyararak,
- Mevcut olan aksonlardaki güçleri değiştirerek.

Aksonların işaretleri değerlendirebilecek yeteneği olduğu savunulur. Aksonda bulunan bu özellik ise bilinen bir sinir için bir işaretin önem derecesinin ne olduğunu vurgulamaktadır (27).

2.3.4.1. Girişler

Yapay sinir hücresine girişler yardımıyla farklı bir yapay sinir hücresinden ya da dış dünyadan bilgi alımı yapılabilir.

2.3.4.2. Ağırlıklar

Yapay sinir hücresinde yer alan ağırlıklar; yardımıyla girişler aracılığıyla hücreye aktarılan bilgilerin önemi ve aynı zamanda bu bilgilerin hücre üzerindeki etkisini ifade eden katsayılardır. Her bir girişe ait ağırlık değeri vardır. Bu ağırlığın değerinin büyüklüğü giriş hakkında hücre için önemli olup olmamasını belirlemekte yeterli olmaz. Ağırlığın artı değer ile ifade edilmesi girişin hücredeki etkisinin pozitif, eksi değer ile ifade edilmesi girişin hücredeki etkisinin negatif olarak belirleneceğini gösterir. Ağırlık değerleri sabit veya değişken değerlerle ifade edilebilirler (27).

2.3.4.3. Toplama İşlevi

Yapay sinir ağındaki toplama işlevi girişler ile o girişlerin sahip olduğu ağırlığın çarpılarak bulunan bu çarpımların toplanmasıdır.

$$Net\ Toplam = \sum_{i=1}^{n} x_i w_i$$

Fakat pek çok uygulamada eşik değeri toplamaya katılmıştır. Eşik değeri θ ile gösterilmektedir ve bu toplamaya aşağıdaki gibi katılmıştır.

$$Net\ Toplam = \sum_{i=1}^{n} x_i w_i + \theta \quad \text{ya da} \quad Net\ Toplam = \sum_{i=1}^{n} x_i w_i - \theta$$

Bu toplama işlemi her modelde ve her uygulamada kullanılmak zorunda değildir. Bazı modellerde kullanılacak olan toplama fonksiyonunu modelin kendisi belirler. Çoğunlukla

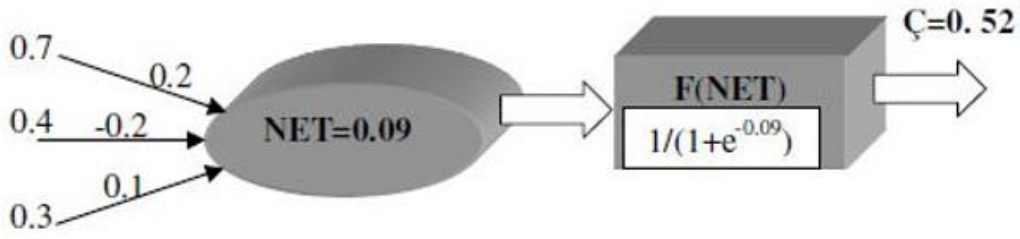
farklı toplama işlemleri kullanılır. Durumların bazılarında girişlerin değeri önemli iken, bazı durumlarda ise sayısı önemli olabilmektedir. Bir problemde kullanılacak olan toplama fonksiyonundan en uygununu belirlemek için bir formül bulunmamaktadır. Bu nedenle problemin amacına uygun olacak toplama fonksiyonunu belirleyebilmek adına deneme yanılma yolu kullanılmaktadır. Ayrıca aynı problemde kullanılan yapay sinir hücrelerinin her birinde aynı toplama işlemi kullanabileceği gibi farklı toplama işlemi de tercih edilebilir (18, 27).

2.3.4.4. Çıkış İşlemi

y = f(v) ile çıkış işlemi gösterilir. Çıkış işlemi aktivasyon fonksiyon sonucunun diğer sinirlere veya dış dünyaya gönderilmesidir.

2.3.5. Yapay Sinir Hücresinin Çalışma Prensibi

YSA hücresi çalışması girişler ve ağırlıklar aşağıdaki şekildeki gibidir.



Şekil 2.6. YSA'nın Çalışmasına Bir Örnek

2.4. YSA MODELLERİ

YSA iki farklı şekilde sınıflandırılır.

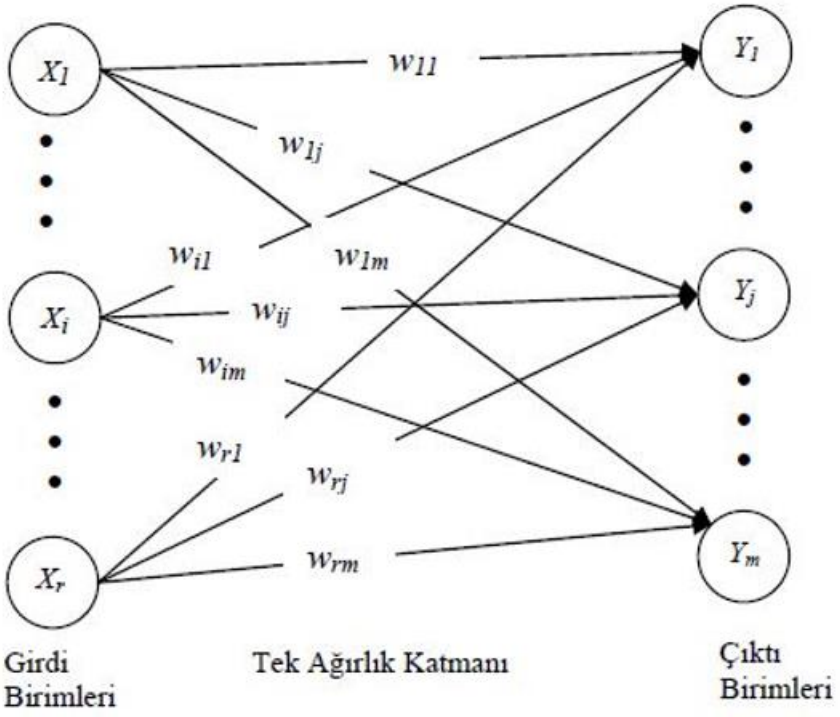
1. Tek katmanlı sinir ağları
2. Çok katmanlı sinir ağları

Katman sayısı belirlenirken, girdi birimi üzerinde işlem yapılmadığından katman olarak sayılmaz (18, 24).

2.4.1. Tek Katmanlı Sinir Ağları

Giriş ve çıkış olmak üzere iki katmandan oluşan YSA modellerine tek katmanlı olan YSA denir. Bu iki katmanda da birden çok giriş değeri ve çıkış değeri vardır. Giriş katmanında yer alan her bir giriş değeri sinaptik bağlantılar ile çıkış katmanına bağlanır. Her bağlantının bir ağırlık değeri vardır. Ayrıca bias sapma değeri ile ağın çıktı değerinin sıfır olmasını engellenir. Tek katmanlı yapay sinir ağları tanıma, örnek ilişkilendirme, örnek sınıflandırma ve buna benzer problemlerin çözülmesinde de kullanılabilmektedir (18, 24).

Aşağıdaki Şekil 2.7' de tek katmanlı olan YSA için bir örnek verilmiştir.



Şekil 2.7. Tek Katmanlı Olan YSA

Tek katmanlı olan YSA modeli eğitiminde kullanılan önemli olan üç metot mevcuttur (18, 24):

1. Hebb Kuralı
2. Perseptron Öğrenme Kuralı
3. Delta Kuralı

2.4.1.1. Hebb Kuralı

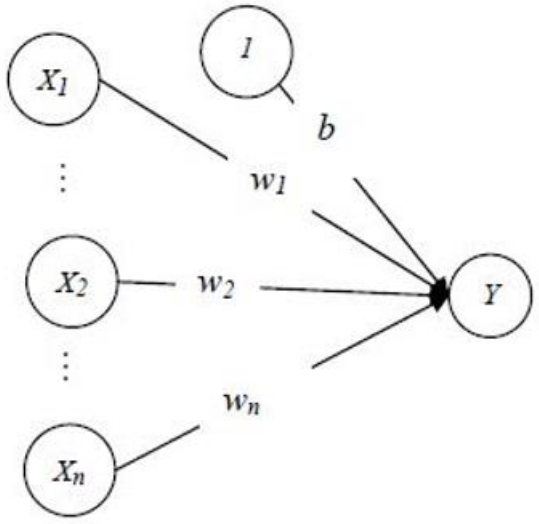
YSA modeli için bilinen en basit en eski olarak kabul edilen öğrenme kuralıdır. Hebb, sinaps ağırlıklarının (uzunluklarını) değişimi ile öğrenmenin gerçekleşeceği önerisinde bulunmuştur. Hebb in önerdiği fikre göre, iki nöronun bağlı olduğu durumda bu iki nörondan her ikisi de "aktif" ise, iki nöron için uygun olacak şekilde ağırlık değerlerinin bir artış göstermesi gerekir. Aynı şekilde, aynı anda iki nöron "pasif" ise, bu durumda ağırlık değerlerinin artırılması gerekmektedir. Bu durum gerçekleştiğinde, öğrenme daha güçlü bir olacaktır (18, 24).

2.4.1.2. Perseptron

YSA modelinin öğrenebilir özelliğine sahip olan ilk bilinen modelidir. Hebb kuralına göre daha üstün yetenekli bir öğrenme kuralıdır. Bu algoritma yakınsama özelliğine ek olarak tekrarlı öğrenme özelliğine de sahiptir. Bahsedilen bu özellikler perseptron modeli için önemli özelliklerin başında gelmektedir (18, 24).

2.4.1.2.1. Perseptron Algoritması

Şekil 2.10' da perseptron algoritmasının şekilsel gösteri verilmektedir. Şekilde Y ile çıktı birimi gösterilirken girdi birimleri X_1 ,..., X_n olarak gösterilmiş olup 1 ile de sapma sinyali gösterilmektedir. w_1 ,..., w_n ağırlıkları, b de sapma değerine ait ağırlığı ifade eder.



Şekil 2.8. Perseptron Mimarisinin Basit Gösterimi

Perseptronlar, YSAnın öğrenebilir niteliğini taşıyan ilk modelidir. Hebb kuralından daha yetenekli bir öğrenme kuralıdır. Perseptron tekrarlı öğrenme algoritmasıdır ve çözümün varlığı durumunda yakınsama niteliğine sahiptir. Bu, perseptron modelinin en önemli niteliklerinden biridir. Rosenblatt (1962) ve Minsky-Papert (1969, 1988) tarafından çeşitli perseptron modelleri tanımlanmıştır. Orijinal perseptronlar, duyumsal birimler, birleştirici birimler ve cevap birimleri olmak üzere nöronların üç durumuna sahiptirler. Örneğin, bir basit perseptron duyumsal ve birleştirici birimler için ikili aktivasyon, cevap birimi için ise +1, 0, veya -1 değerlerini üreten aktivasyon uygulayabilir. Sınıflandırma problemlerinde eşik değerli aktivasyon fonksiyonu kullanılır (18, 24) :

$$f(y\_in) = \begin{cases} -1, & y\_in < -\theta \text{ ise} \\ 0, & -\theta \le y\_in \le \theta \text{ ise} \\ 1, & y\_in > \theta \text{ ise} \end{cases}$$

Çıktı biriminin aktivasyonu y= f(y _ in) şeklinde hesaplanır.

Birleştirici birimden cevap birimine giden bağlantıların ağırlıkları perseptron öğrenme kuralı ile ayarlanır. Her eğitim girişi için, sinir ağı, çıkış biriminin cevabını hesaplar. Daha sonra sinir ağı, bu örnek için çıkış değeri ile hedeflenen çıkış arasındaki farkı karşılaştırarak bir hata oluşup oluşmadığını tespit eder. Yapay sinir ağı, hesaplanmış çıkış değeri "0" ve hedef değeri "-1" olan örnek için hatayı ayırt edemez, buna karşıt olarak hesaplanmış çıkış değeri "+1" ve hedef değeri "-1" olan örnek için hatayı ayırt edebilir. Bu durumlarda, hedef verinin işareti yönünde ağırlıkların işareti değiştirilmelidir. Bununla birlikte çıkış birimine "0" olmayan sinyaller gönderen bağlantıların ağırlıkları ayarlanmalıdır. Eğer belirli bir eğitim giriş örneğinde hata oluşuyorsa, ağırlıklar

$$w_i(yeni) = w_i(eski) + \alpha t x_i$$

formülüne göre değiştirilmelidir.

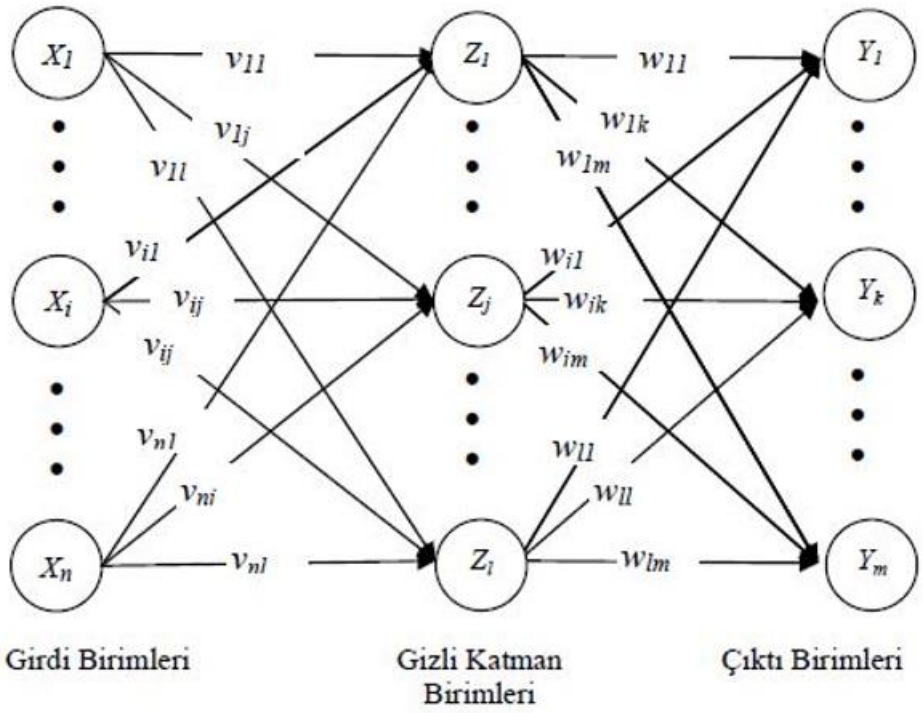
Burada hedef değeri t ya "+1" ya da "-1"dir ve α öğrenme oranı katsayısıdır. Eğer hata oluşmadıysa ağırlıklar değiştirilmemelidir. Eğitim işlemi hata oluşmayıncaya kadar devam etmelidir. Bu kuralın amacı, ağın tam olarak doğru cevap veremediği eğitim örnekleri için ağırlıkları ayarlamaktır. Ayrıca, eğitim sonunda bu ağ sınırsız sayıdaki eğitim adımları için ağırlıkların değerlerini bulmalıdır (18, 24).

2.4.1.3. Delta

Delta modelinde bütün girdi değerleri için hedef ve çıktı farklarının karesinin toplamının, yani, toplam hatasını küçültülmek asıl hedeftir. Delta kuralında amaç, bütün eğitim numunelerindeki hataları en aza indirgemektir.

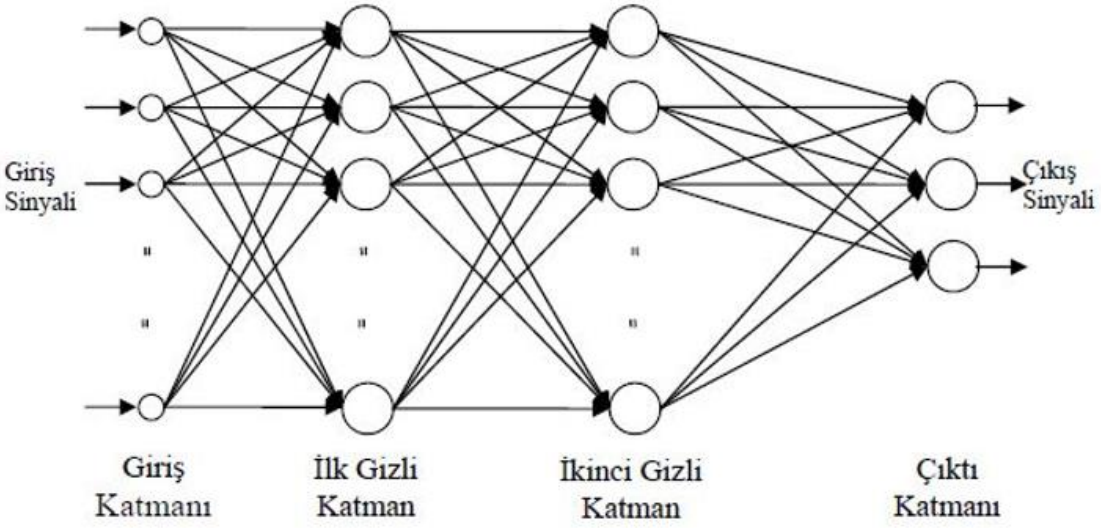
2.4.2. Çok Katmanlı Sinir Ağları

Ayrılamayan problemlere çözüm bulmada tek katmanlı ağlar başarısız olmaktadırlar. Bu sebeple bilim adamları çok katmanlı olan YSA modellerini incelemişlerdir. Çok katmanlı olan yapay sinir ağları modellerinde eğitim algoritması geliştirmek en önemli olan aşamalardan biriydi. Hinton, Williams ve Rumelhart tarafından 1986 yılında bu gerçekleştirildi. Bu eğitim yöntemi standart geriye yayılım (backpropagation) olarak isimlendirilir. Geriye yayılım metodu yardımıyla hata kareler toplamının küçültülmesi düşüncesine dayanmaktadır ve bu delta kuralının genelleştirilmiş hali kullanmaktadır. Dolayısıyla bu metod her adımda hatayı küçültülmek için, Widrow-Hoff eğitiminde olduğu gibi, gradient azalış metodu kullanılmaktadır. Bu durum söz konusu olduğunda ise doğrusal olmayan aktivasyon fonksiyonları gizli olan katmanda uygulanmaktadır. Örnek olarak lojistik sigmoid fonksiyonu ve ona uygun olacak şekilde delta kuralının genelleştirilmiş hali uygulanmaktadır (26, 27). Şekil 2.9'da çok katmanlı yapay sinir ağları için örnek gösterilmiştir.



Şekil 2.9. Tek Gizli Katmanı Olan Çok Katmanlı YSA

Çok katmanlı olan algılayıcılar (ÇKA), bilgi girişinin olduğu girdi katmanını, bir ya da birden çok sinir hücresini içeren gizli katmanlarını ve çıktı katmanını içermektedir (25). Ağ boyunca girdi katmanından gelen sinyaller ileri yönde katmandan bir diğer katmana olacak şekilde yayılırlar (Şekil 2.10).



Şekil 2.10. Çok Katmanlı YSA

Ağda bulunan farklı katmanlar boyunca iki tane geçiş vardır. Bunlar geri yayılım ve ileri yayılım olarak isimlendirilir. İleri yayılım algoritmasında ağdaki girdi vektörü giriş katmanına uygulanır. Girdinin etkisi bir katmandan diğer bir katmana ağ boyunca yayılır. İleri yayılım algoritmasında ağda bulunan sinaptik ağırlıklar bellidir. Geri yayılım algoritmasında ise bu sinaptik ağırlıklar hepsi uyumlu olacak şekilde bir hata-düzeltme kuralıyla düzenlenir. Ağdaki gerçek bir çıktı isteğe bağlı olarak bir çıktıdan çıkartılarak bir hata sinyali üretilir. Bu hata sinyali ise ağda sinaptik bağlantılarla ters yönde olacak şekilde geriye doğru olacak şekilde yayılır. Sinaptik ağırlıklar ağda elde edilen gerçek çıktıyı istatistiksel anlamda istenen çıktıya yakın olması için düzenlenir (3, 9, 11).

2.5. Karar Ağaçları

Karar Ağaçları, verilerin belirli özelliklerini göz önüne alarak sınıflandırma yapmayı amaçlar. Bunu yapabilmek için algoritmada çeşitli özelliklere sahip olan girdi belirlenir. Çıktıyı da belirli bir veri özelliğine göre seçerek algoritmanın bu çıktının sahip olduğu özelliğe ulaşmak için girdi özelliklerinden hangilerinin bir arada bulunması gerektiğini keşfederek bunu ağaç veri yapısı şeklinde göstermesi sağlanır (28).

Bu teknik uygulanırken sınıflandırma yapabilmek için öncelikle bir ağaç yapısı oluşturulur, daha sonra ise bu ağaca veri tabanında bulunan kayıtların her biri uygulanır ve kayıt bulunan sonuca göre sınıflandırılır. Karar ağaçları veri çok karışık olsa bile, bağımsız değişkenleri ve bağımsız değişkenlerin model için önemini görsel olarak kolayca anlaşılır bir ağaç yapısı biçimi ile sunmaktadır (10, 28).

Karar ağacı yönteminde verinin sınıflandırılması iki adım ile gerçekleşmektedir. İlk adımda model oluşturmak için önceden bilinmekte olan bir eğitim verisinin sınıflama algoritması ile çözümlendiği öğrenmenin gerçekleştiği basamak yer alır. Öğrenilmiş olan model, karar ağacı veya sınıflandırma kuralları ile verilir. Sonraki adım ise eğitim verisinin karar ağacının veya sınıflandırma kurallarının doğruluğunu kestirmek için test edilerek sınıflandırılmanın yapılmasıdır (10).

Karar ağacı algoritmasını bir probleme uygularken aşağıdaki şartların sağlanması gerekmektedir:

- Nesnelerin belirli sayıda özelliklerle ifade edilebilir olması gerekir. Örneğin; sıcak, soğuk vb. özelliklerin kesikli veya sürekli olması fark etmez.
- Sınıfları belirlerken ayırt edici özellikler olmalıdır. Karar ağacı dallar, karar düğümleri ve yapraklardan oluşur (29).

Karar ağaçları ağaca benzer bir yapıda olup kökten, dallardan ve yapraklardan oluşur, verideki tüm gözlemleri kapsayacak olan kök ile başladıktan sonra aşağılara inildikçe veriyi alt gruplarına ayıracak biçimde dallara bölünür. Kökten başlayıp dallar boyunca büyümekte olan ağaç yapısındaki her bir boğuma "düğüm" denir, elde edilen ağaçlardaki birbirine benzer olan düğümlere "terminal düğüm (parent node)", benzemeyen düğümlere ise "yavru düğümü (child node)" adı verilir (18, 30). Düğümler üzerinde niteliklerine göre test edilip bu işleminin sonucuna göre de ağacın veri kaybına uğramadan dallara ayrılması sağlanır. Her düğümde

ardışık olarak test işlemi ve dallara ayrılma işlemi gerçekleşmekte ve sonunda ağaç sınıflarla son bulacaktır (30).

- **Düğüm:** Veriye uygulanacak olan test belirlenir. Düğümlerin her biri tek özellikteki testi göstermektedir. Testin sonucuna göre ağacın dalları meydana gelir. Veri kaybı olmaması için dalları oluştururken verilerin tamamını kapsayacak şekilde farklı dallar oluşturulmalıdır.
- **Dal:** Test işleminin sonucunu gösterir. Elde edilmiş olan her dal ile sınıf belirlenir. Ancak sınıflandırma yapılamıyorsa yeniden bir karar düğümü meydana gelir. Karar düğümü sonucunda elde edilmiş olan dallardan sınıflandırmanın tamamlanıp tamamlanmadığını tekrardan kontrol ederek devam edilir.
- **Yaprak:** Daldan sonra sınıflandırma işlemi yapılabiliyorsa yaprak oluşumu meydana gelir. Veriler kullanılarak elde edilen her bir sınıftan birini oluşan yapraklardan biri tanımlar (30).

Karar ağacı modelinde, tanımlanan soruya verilen cevap gruplara bölünmektedir. Bir soruya göre grup meydana geldikten sonra ve gruplar arasında var olan risk maksimize edilip oluşmuş olan iki tane grup için işlemlere devam edilir. İstatistiksel olarak farklılık bulununcaya kadar işlemlere devam edilir, istatistiksel anlamda bir farklılık bulunmazsa işlemler durdurulur (31).

Karar ağacı modeli oluşturulurken bahsedilen ağaçları oluşturma ve budama aşamaları vardır.

2.5.1. Ağaç oluşturma

Veri kaynağında bulunan tüm nesneleri içine alan kök düğümünden başlar, tekrarlamalı bir şekilde her bir düğümde yer alan nesneleri seçilen bir özelliğe göre farklı dallara bölerek tüm nesneleri sınıflandırarak yaprak düğümlere bölünene kadar, ya da ayırt edici bir özellik kalmayıncaya kadar bölünme devam eder. Nesneleri alt düğümlerine ayırırken alt düğüme alınacak olan nesneler ne kadar birbirine benzer özellik gösterirse düğümde yapılacak dallanmalar da bir o kadar kuvvetli olacaktır. Bu nedenle, düğümlerde sınaması yapılan özellik homojenlik açısından en yüksek kazanç sağlayacak olan özellik olarak belirlenir (10, 31).

2.5.2. Ağaç Budama

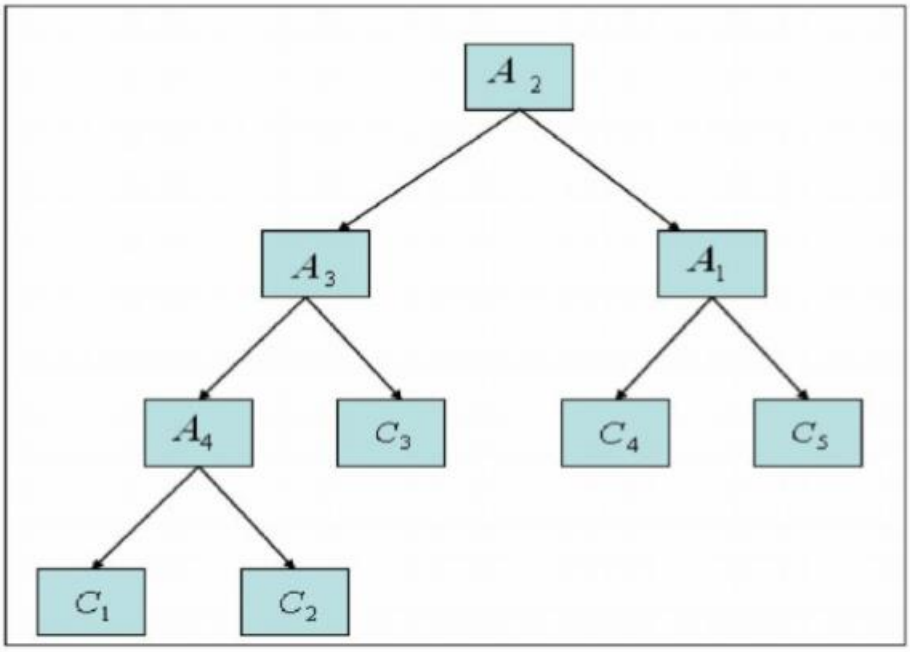
Bir önceki basamakta, verilerin tamamı karşılaştırılacak özellik kalmayıncaya kadar bölünür ya da aynı sınıftaki üyelerden oluşacak olan yapraklara bölününceye kadar devam eder. Bu algoritmanın sonunda çok fazla derin ya da çok daha az olan deneme kümesi örneğini kapsayan yaprak düğümlere sahip olan ağaçlar meydana gelebilir. Öğrenme kümesinin üzerinde böyle bir ağaç test edildiğinde doğruluğu çok yüksek olan sonuçlar verir. Fakat şimdiye kadar görülmeyen örneklerle karşılaştığında bu model çok kötü doğruluğa sahip olan sonuçlar meydana getirebilir. Bu şekilde olan model verimli olamamakla beraber veriyi de genelleyemez. Böyle olan bir modele aşırı uyum (overfitting) özelliğine sahiptir denir. Bu özelliğe sahip olmak model için istenilmeyen bir durumdur (31, 32).

Aşırı uyum durum varlığı genellikle veride bulunan gürültüden (yanlış değerli değişkenler, hatalı sınıf değeri) kaynaklanabileceği gibi rastgelelikten ya da problem alanının karmaşık oluşundan kaynaklanabilir. Aşırı uyum durumunu azaltabilmek için ağaçlarda yapılacak olan işlem budama işlemidir. Budama işleminde, bazı dalların veya alt dalların kaldırılıp kaldırılan dala ait olan nesnelerin yoğun bir şekilde bulunduğu sınıfla etiketlenen yaprak düğümlere yerleştirilmesi sonucunda gerçekleştirilir. Ağaç oluşturulma aşamasında erken-dur metoduyla erken-budama gerçekleşebileceği gibi ağaç oluşumu tamamlandıktan sonra geç-budama da gerçekleşebilir. Budama metotlarından geç-budama metodunun daha başarılı bir metot olduğu bilinmektedir. Geç-budama yönteminde ağaç zaten oluştuğu için dalların hangisinin kullanılmayacağı ve aşırı uyumun söz konusu olduğu görülmektedir. Geç-budama yöntemi uygulanırken düğümlerde oluşan beklenen hata miktarı göz önüne alınır. Bir düğümde bakılan beklenen hata değeri, o düğümün sahip olduğu alt dallarda oluşan beklenen hata değerinden daha küçük olması durumunda alt dallar budanır (31, 33).

Veri tabanı $D = \{t_1,...,t_n\}$ olsun. t_t, $t_i = \{t_1...t_{in}\}$ den ve bu D veri tabanı da $\{A_1,A_2,...,A_n\}$ alanlarından oluşsun.

Bunların dışında sınıflar da $C = \{C_1,...,C_n\}$ ile verilmiş olduğunda,

- ✓ Düğümlerin her biri A_t alanı ile gösterilmiş,
- ✓ Düğümlerin her birinden ayrılmış kollar bu alanla ait olan bir soruya cevap veren,
- ✓ Yaprakların her birinin bir sınıf olarak gösterildiği karar ağacı şekil 2.11'de verilmiştir (32).



Şekil 2.11. Karar Ağacı Modelinin Yapı Olarak Gösterimi

Şekil 2.11'de verilen karar ağacı modelinde düğümlerin her biri A_1, A_2 ,..., A_n'ler ile gösterilmekte ve her bir düğüm iki tane dala ayrılmaktadır. Düğümden dala ayrılma işlemi yapılırken, veri tabanında A_i düğümü ile ilgili cevabı bulunacak olan bir soru sorulup verilmiş cevaba göre de bir dal takip edilmektedir. Ağaçtaki yaprakların her biri C_1 ,C_2 ,...,C_n 'ler ile gösterilmektedir. Aynı zamanda C_1 ,C_2 ,...,C_n 'lerin her biri bir sınıfı da temsil etmektedirler. Karar ağaçlarının oluşumunda önemli bir husus kullanılacak algoritmanın ne olduğudur. Çünkü ağacın şekli kullanılacak olan algoritmaya değişebilir. Bu nedenle kullanılan algoritmaya göre değişik şekilde ağaç yapıları oluşmakta ve farklı sınıflandırma sonuçları ortaya çıkmaktadır. Kök olarak da bilinen ve ilk düğümü meydana getiren A_i 'nin farklı olması, son uçtaki yaprağa giderken takip edilecek yolu değiştirecek ve sonuç olarak sınıflandırma da değişecektir (18, 32).

Değişken seçimi esnasında tekrarlamalı algoritma karar ağacı modelindeki döngüden çıkabilmesi için o düğümde bulunan bütün elemanların aynı sınıfta olması şartı bulunmaktadır. Eğer kalmış olan değerler yalnızca aynı sınıfta bulunuyorsa ya da sınıflandırma yapılacak değer yoksa karar ağacı modelindeki döngü son bulur ve karar ağacı modeli oluşumu tamamlanmış olur. Sonuç olarak meydana gelen sınıflardaki elemanların her birinin özelliği aynı sınıfta bulunan diğer elemanların özelliklerine benzerlik gösterir. Ağaç

yapısı ile heterojen yapıda bulunan veri kümesinden homojen ve daha küçük yapıya dönüşebilmesi için koşullar tanımlamaktadır. Ağaç oluşumunun son aşamasında oluşan ağaca maksimum ağaç denir. Oluşan bu maksimum ağacın iki dezavantajı vardır (33).

- ✓ Eklenmiş olan her bir bağımsız değişken için maksimum ağaç hatalı sınıflama oranını düşürdüğünden başlangıçta kullanılan veri setini mükemmel biçimde tanımlar. Böyle bir durum olduğunda, maksimum ağaç kullanılan veri için daha iyi olan bir tahmin modeli oluşturur. Fakat, başlangıçtaki veri seti üzerinde aşırı uyumlu olan maksimum ağaçlar farklı olan bir veri seti için iyi bir tahmin sağlayamazlar.
- ✓ Bir sınıflama ağacında terminal düğüm sayısı karmaşıklık ölçüsüdür. Terminal düğümün sayıca fazla olması yani karmaşıklık ölçüsü yüksek bir maksimum ağacın hem yorumlanması zordur hem de anlaşılması güçtür.

Oluşan böyle sorunların çözümlenebilmesi amacıyla maksimum ağacı budamak gerekir. Maksimum ağaç budandığında küçük olan ağaçlar dizini oluşur ve oluşmuş olan bu dizinde yer alan ağaçlar içerisinden optimum ağacın seçimi yapılır. Seçilen optimum ağaç oluşturulan maksimum ağaca göre daha sade bir yapıya sahip olup başlangıç veri setine oluşturulan maksimum ağaçtan uyum olarak daha az uyuma sahip olup sınıflandırmadaki hatası daha yüksek olacaktır (30-32).

"Karar ağacı yönteminde verinin sınıflandırması üç aşamada gerçekleşir."

- **Öğrenme:** Sonuçları daha önceden bilinen eğitim verisinden model oluşturulur.
- **Sınıflama:** Modele yeni bir test verisi uygulanarak karar ağacının doğruluk oranı belirlenir. Test verisi uygulanmış olan modelin doğruluğu, yapmış olduğu doğru sınıflandırmanın test verisinde bulunan bütün sınıflara oranıdır.
- **Uygulama:** Doğruluk oranı kabul edilen oranda ise, yeni verilerin sınıflanmasında karar ağacı modeli uygulanır (18).

Karar ağacı modeli oluşturulurken test verisini en iyi biçimde dallara ayıran özellik tespit edilir ve tespit edilen bu özellik daha önce seçilir. Bu durumda daha da iyi bir karar ağacı modeli elde edilir. En iyi biçimde dalların ayrılmasında kullanılan özelliğin tespitinde çeşitli ölçütler vardır.

Aşağıda bahsedilen çeşitli ölçütlerden bazıları verilmiştir:

i' sınıfına ait olan verilerin, verilmiş olan *t* düğümündeki bölünmesi $p(i|t)$ ile gösterilsin.

c ile de sınıf sayısı gösterilsin ve entropi hesaplanırken $0 \log_2 0 = 0$ olarak düşünülmüştür.

$$\text{Entropi(t)} = -\sum_{i=0}^{c-1} p(i|t) \log_2 p(i|t),$$

Bir düğümün ne derece bilgilendirici olduğunu ölçerken entropi kullanılır. Bununla "İyi" ile ne denilmek istendiği belirtilir.

$$Gini(t) = 1 - \sum_{i=0}^{c-1} [p(i|t)]^2 ,$$

$$Sınıflandırma\ hatası(t) = 1 - max_i[p(i|t)],$$

Karar ağacı modeli meydana geldikten sonra, kurallar kök ile başlayıp yaprağa doğru ilerleyerek belirlenir. Fakat fazla miktarda dal ve düğümden meydana gelen karar ağacı modelinde kuralları belirlemek zor hale gelir. Karar ağacı modeli daha okunabilir olması için kuralların yazarken şartlı ifade olan IF-THEN (Eğer-ise-O Zaman) kullanılır. IF (Eğer ise) ifadesi olan kısım daldan sonuna kadar olan yoldaki bütün testleri kapsarken THEN (O Zaman) ifadesi olan kısım ise son sınıflamayı göstermektedir. IF THEN yapısında bulunan kurallara da Karar Kuralları denir (13).

Karar ağaçları modeli, model kurulumunun basit olması, kolay bir şekilde veritabanı sistemleriyle entegre olması, sonuçların daha kolay bir şekilde yorumlanması ve güvenilirliklerinin de iyi olması nedenlerinden dolayı sınıflama modellerinin içinde yaygın olarak kullanılan bir yöntemdir (34).

Karar ağaçlarının diğer yöntemlere göre güçlü olan özellikleri özet olarak aşağıdaki şekildedir:

- ✓ Kurallar daha anlaşılabilir şekilde üretilir.
- ✓ Sınıflandırmayı aşırı hesaplama yapmadan yaparlar.
- ✓ Hem kesikli değişkenler hem de sürekli değişkenler için uygundur.
- ✓ Sınıflandırmada ve kestirimde hangi alanların en önemli olduğunu açık biçimde gösterir (35).

2.6. Diyabet Nedir?

Şeker Hastalığı, diğer bir adı ile Diyabet; pankreasın üretmiş olduğu insülin hormonunun eksikliği, etkisizliği ya da yokluğundan kaynaklanan yaşam boyu sürmekte olan kan şekerinin yüksekliği ile karakterize edilmiş bir metabolizma hastalığıdır. Günümüzde sıkça görülen diyabet, tıp dilinde ise "diabetes mellitus" şeklinde ifade edilmektedir (36).

Normal koşullar altında besinlerden elde edilen şeker yani glikoz pankreasın ürettiği hormon olan insülin sayesinde hücrelerden içeri girdikten sonra enerji ihtiyacını gidermek için kullanılır. Hücre, enerji ihtiyacının %90'ından fazlasını glikozdan elde eder. İnsülin eksikliği ya da insülin yokluğu sonucunda kanda bulunan şeker hücre içerisine alınamaz ve kandaki şeker yükselir. Kan şekerinin yükselmesi sonucunda ağız kuruması, kilo kaybetme, çabuk yorulma, suyu çok içme, idrara sık çıkma, halsizlik gibi bulgular görülür. Şeker hastalığı, her yaşta görülme ihtimali olan, insanı doğrudan etkileyen, yaşam boyu devam eden ve iyi bir şekilde tedavisi yapılmadığında sağlık yönünden önemli olan başka sorunlara neden olan, fakat iyi bir şekilde tedavisi yapıldığında, uzun ve sağlıklı bir yaşam sürdürme imkanı olan bir hastalık türüdür. Şeker hastalığı tüm dünyada hızlı bir şekilde yayıldığı gibi Türkiye' de de hızlı bir şekilde yayılmakta olan çok önemli sağlık problemidir.

Tip 1 ve Tip 2 olmak üzere diyabet hastalığının iki tipi vardır. Hasta olan kişinin vücudunda insülin üretilmiyorsa Tip 1; hastanın vücudunda üretiliyor fakat kullanılmıyorsa da Tip 2 olarak adlandırılır. Bu tezin amacına yönelik olarak veriler Tip 2 diyabet hastalarından toplandı. Tip 2 diyabetin meydana gelmesinde önemli iki mekanizma vardır. Birincisi pankreasın üretmiş olduğu insülinin hücreden içeriye alınamaması, ikincisinde ise pankreasın üretmiş olduğu insülinin azalması. Tip 2 diyabette insülinin hücreden içeriye alınamaması insülin direnci olarak bilinir. İnsülinin hücreden içeri alınamamasının sonunda ise hücrede enerjinin kaynağı olarak bilinen glikoz hücreden içeri alınamaz. Bunun sonucunda kanda birikip kan şekerini yükseltir. Daha sonra ise pankreasın üretmiş olduğu insülin miktarı giderek azalmaya başlar ve sonuç olarak diyabet daha da ilerler (36).

2.6.1. Diyabet Tanısı

Diyabet tanısı için kullanılmakta olan iki temel test vardır. Bunlar; açlık kan şekerinin ölçümü ve Oral Glikoz Tolerans Testi (OGTT) yani şeker yükleme testidir. Sağlıklı bireyler için açlık kan şekeri seviyesi ortalama olarak 70-100 mg/Dl arasındadır. Diyabet tanısına karar verilebilmesi için açlık kan şekerinin 126 mg/Dl'nin yukarısında olması yeterlidir. Bu değer 100-126 mg/Dl arasında olduğunda bireyin tokluk kan şekeri araştırılmalıdır bunun için

OGTT uygulanır. Yemek yemeye başlanan saatten iki saat sonra kan şekeri ölçülür ve bunun sonucunda kandaki glikoz seviyesi 200 mg/Dl'nin yukarısında olursa diyabet hastalığı göstergesi iken, 140-199 mg/Dl arasında olursa gizli şeker göstergesidir. Bunlara ek olarak HbA1C testinin yani son üç aylık kan şekerinin %7'nin üzerinde olması da diyabet tanısının göstergesidir (36).

3. MATERYAL VE METOT

3.1. Çalışma İzni

Bu çalışma, Malatya Klinik Araştırmalar Etik Kurulu'nun 2016/144 protokol numaralı izni ile onaylanmıştır (Ek-2).

3.2. Çalışmada Kullanılan Veri Seti

Çalışmada, İnönü Üniversitesi Tıp Fakültesi Turgut Özal Tıp Merkezi İç Hastalıkları Anabilim Dalı Diyabet ve Tiroid polikliniğine gelen Tip 2 Diyabet Mellitus olan ve olmayan hastalardan toplanan veriler kullanılmıştır. İlgili veri seti, 146'sı (%46.6) Tip 2 DM bulunan, 167'Si (%53.4) Tip 2 DM bulunmayan toplam 313 hasta kayıtlarından oluşmaktadır. Bu kapsamda, Tip 2 DM ile ilişkili olabilecek olası risk faktörlerine ilişkin tanımlayıcı tablo Tablo 3.1'de gösterilmiştir.

Tablo 3.1. Tip 2 DM ile ilişkili olabilecek olası risk faktörlerine ilişkin tanımlayıcı tablo

Değişkenler	Veri Tipi	Açıklama	Değişken Türü
Tip 2 Diyabet Mellitus(DM)	Nitel	Var/Yok	Bağımlı
Cinsiyet	Nitel	Kadın/Erkek	Bağımsız
Aile Öyküsü	Nitel	Var/Yok	Bağımsız
Uzun süre ilaç kullanımı	Nitel	Var/Yok	Bağımsız
Kortizon Kullanımı	Nitel	Var/Yok	Bağımsız
Eşlik eden hastalık	Nitel	Var/Yok	Bağımsız
Yüksek tansiyon	Nitel	Var/Yok	Bağımsız
Stres faktörü	Nitel	Çok var/Az var/Yok	Bağımsız
Kalp hastalığı	Nitel	Var/Yok	Bağımsız
Kolestrol yüksekliği	Nitel	Var/Yok	Bağımsız
Sigara kullanımı	Nitel	İçiyor/İçmiyor	Bağımsız
Alkol tüketimi	Nitel	Var/Yok	Bağımsız
Egzersiz durumu	Nitel	Düzenli /Düzensiz/Arasıra	Bağımsız
Karbonhidrat Kullanımı	Nitel	Var/Yok	Bağımsız
Sebze Kullanımı	Nitel	Var/Yok	Bağımsız
Et kullanımı	Nitel	Var/Yok	Bağımsız
Yaş	Nicel	Doğal sayı	Bağımsız
Kilo	Nicel	Doğal sayı	Bağımsız
Boy	Nicel	Doğal sayı	Bağımsız
Başlama yaşı	Nicel	Doğal sayı	Bağımsız
Günlük ekmek tüketimi	Nicel	Pozitif reel sayı	Bağımsız
Yüksek yoğunluklu lipoprotein (HDL)	Nicel	Pozitif reel sayı	Bağımsız
Düşük yoğunluklu lipoprotein (LDL)	Nicel	Pozitif reel sayı	Bağımsız
Trigliserid	Nicel	Pozitif reel sayı	Bağımsız
Total Kolesterol	Nicel	Pozitif reel sayı	Bağımsız
Açlık Kan Şekeri	Nicel	Pozitif reel sayı	Bağımsız

3.3. Örneklem Büyüklüğü

Bazı kaynaklarda (12) çok değişkenli istatistiksel modellerde n olarak bilinen gözlem sayısı yani örneklem sayısı çalışmada bulunan bağımsız değişkenlerin sayısının en az 5 katı olmalıdır

$$n > 5k \ (k: bağımsız\ değişken\ sayısı)$$

koşulunun kullanılabileceği belirtilmiştir. Çalışmadaki veri setinde 146'sı Tip 2 DM'li, 167'si Tip 2 DM'li olmayan toplam 313 hastadan oluşmaktadır. k=25 olduğu için, hem Tip 2 DM'li olan ve TİP 2 DM'li olmayan hasta sayısında yukarıdaki koşulun sağlandığı görülmüştür.

3.4. Kullanılan Yöntemler

Çalışmanın yapılacağı veri setinde açlık kan şekeri, kolestrol yüksekliği, HDL, LDL ve günlük ekmek tüketimi bağımsız değişkenlerinde veri setinde kayıp değerler yer almaktadır. Kayıp değerlerle ilgili problemlerin giderilmesi için kayıp değer atama yöntemlerinden biri olan k-En Yakın Komşu algoritması (kNN) kullanıldı. Bu atama işlemi ise Rapidminer Studio Free 8.1.000 versiyonu (37) ile yapıldı. kNN tabanlı algoritmada atama yapılırken, kayıp olan gözlem değerlerine benzeyen (benzerlik ölçütü ise genel olarak Öklidyen uzaklık seçilir.) diğer gözlem değerleri ele alınarak yapılır. Kayıp değer olan bağımsız değişken sürekli sayısal ise, en yakın gözlemlerden k tanesinin ağırlıklı ortalamasıyla değiştirilir. Burada bulunan ağırlık değerleri için Öklidyen uzaklık değerleri tersi alınır (38, 39). Tip 2 DM'li olan ve olmayan hastalara ait veriler kullanılarak çok değişkenli istatistiksel yöntemlerden olan Yapay Sinir Ağları (YSA), Lojistik Regresyon Analizi (LRA) ve Karar ağaçları yöntemleri karşılaştırılmıştır. Bu yöntemler uygulanırken öncelikle veri seti %70'i eğitim veri seti ve %30'u test veri seti olarak ayrılmıştır.

Parametre optimizasyonu modellerin tahmin performansını etkileyebilecek önemli etkenlerden birisidir. Her model tahmin performansını etkileyecek belirli parametrelere sahiptir. Veri setinden daha doğru sonuçları elde edebilmek, modelin performans çıktısının gücünü arttırabilmek için modelin sahip olduğu parametrelerin çok iyi şekilde ayarlanması gerekmektedir. Destek vektör makinesi, yapay sinir ağı gibi modellerde tahmin performansı parametre değerlerinin ayarlanması ile doğrudan ilişkilidir (40, 41). Bu nedenle RapidMiner Studio yazılımı, otomatik olarak en iyi parametre değerlerinin belirlenmesini sağlar. RapidMiner Studio yazılımında parametre optimizasyonu işlemini yapan üç yöntem vardır. Bu yöntemler Grid (Izgara), Quadratic (ikinci derece) ve Evolutionary (Evrimsel) algoritmalarıdır. Bu çalışmada Yapay Sinir Ağları modelinin parametrelerini optimize etmek

amacıyla Evolutionary (Evrimsel) algoritması parametre optimizasyon yöntemi kullanılmıştır. Bu operatör, Grid (Izgara) ve Quadratic (ikinci derece) operatörlerinden daha uygun olan evrimsel bir yaklaşım kullanarak bir dizi parametre için en uygun değerleri bulur ve daha iyi tahmin sonuçlarını hesaplar. Evrimsel operatörü bir alt işleme sahip olup iç içe geçmiş bir operatördür. Belirtilen parametreler için en uygun değerleri bulmak amacıyla alt işlemlerini birkaç kez tekrarlar ve ayrıca model parametrelerini optimize edebilmek için aralıklarla ilgili bir bilgi olmasa da sonuç almayı sağlar (37, 41).

İlgili sınıflandırma modelleri eğitim veri setinde eğitilirken test veri setinde ise öğrenme performansları incelenmektedir. Modellerin hem eğitim hem de test performansı sonuçları ayrı ayrı incelenmiştir. Yöntemlerin sınıflandırma performanslarını karşılaştırılırken performans ölçütlerinden olan doğruluk, kesinlik, sınıflama hatası, AUC (ROC eğrisi altında kalan alan), duyarlılık, seçicilik ve F- ölçümü kullanılmıştır.

- Doğruluk: Doğru sınıflandırılmış örnek sayısının, toplam örnek sayısına oranıdır (24).
- Kesinlik: Pozitif örnek sayısının, model tarafından pozitif tanısı koyulmuş toplam örnek sayısına oranıdır (42).
- Sınıflama hatası: Yanlış sınıflandırılmış örneklerin sayısı veya başka bir deyişle yanlış tahminlerin yüzdesi (37).
- Duyarlılık: Testin, gerçek hastalar içinden hastaları ayırma yeteneğidir (43).

  Tablo 2.1' den

$$Duyarlılık = \frac{a}{a + c}$$

- Seçicilik: Testin, gerçek sağlamlar içinden sağlamları ayırma yeteneğidir (43).

$$Seçicilik = \frac{d}{d + b}$$

- AUC: Duyarlılık y ekseninde, (1-seçicilik) ise x ekseninde olmak üzere oluşturulan saçılım grafiğindeki noktalar bir eğri ile birleştirilir ve bunun sonucunda ROC (Reciever Operator Characteristics, Alıcı İşletim Karakteristiği) eğrisi oluşur. ROC eğrisi altında kalan alanın değeri AUC (Area Under Curve) olarak belirtilir. Bu değer 1'e yaklaştıkça sınıflandırma performansı artar.
- F-ölçümü: Duyarlılık ve kesinlik ölçütlerini birlikte değerlendirebilmek amacıyla bu iki değerin harmonik ortalaması alınarak bulunur (43).

Veri setleri kullanılarak çok değişkenli istatistiksel yöntemlerin analizleri Rapidminer Studio Free 8.1.000 versiyonu (37) ile yapılmıştır.

4. BULGULAR

Çalışmada kullanılmış olan veri seti 25 bağımsız değişken ve 1 bağımlı değişken oluşmaktadır. Bağımlı değişken olan Tip 2 DM değişkeninin dağılım tablosu Tablo 4.1'de gösterilmiştir.

Tablo 4.1. Tip 2 DM değişkeninin dağılım tablosu

Tip 2 Diyabet Mellitus			
Var		**Yok**	
Sayı	**Yüzde**	**Sayı**	**Yüzde**
146	46.6	167	53.4

26 değişken içerisinden kayıp değer içeren değişkenler Tablo 4.2'de gösterilmiştir.

Tablo 4.2. Değişken bazında kayıp değer sayıları

Değişkenler	**Günlük Ekmek Miktarı**	**HDL**	**LDL**	**Trigliserid**	**Kolesterol Yüksekliği**	**Açlık Kan Şekeri**
Kayıp değer	1	1	1	1	2	22

Kayıp değer olan gözlemlere, değer ataması k-En Yakın Komşu algoritması (kNN) algoritması kullanılarak yapılmıştır. Kayıp değer olan gözlemlere atama yapıldıktan sonra tez verisinde bulunan bağımsız değişkenlerden sayısal olanların tanımlayıcı istatistik tablosu ayrıntılı olarak Tablo 4.3'de gösterilmiştir.

Tablo 4.3. Araştırmadaki nicel bağımsız değişkenlere ilişkin tanımlayıcı istatistiksel ölçütlerin dağılımı

	Tip 2 Diyabet Mellitus: Var n=146		**Tip 2 Diyabet Mellitus: Yok n=167**	
Değişkenler	$\bar{X}$**±SD**	**Ortanca (Min-Maks)**	$\bar{X}$**±SD**	**Ortanca (Min-Maks)**
Yaş	54.5 ± 13.2	57.0 (19.0-80.0)	44.2 ± 12.8	43.0 (19.0-76.0)
Kilo	81.2 ± 15.7	79.0 (50.0-135.0)	75.6 ± 14.8	75.0 (45.0-169.0)
Boy	166.2 ± 8.9	165.0 (148.0-190.0)	162.0 ± 21.3	165.0 (1.0-190.0)
Başlama yaşı	45.8 ± 13.3	47.0 (8.0-78.0)	38.0 ± 13.8	38.0 (0.0-72.0)
Günlük Ekmek Miktarı	0.96 ± 0.62	1.00 (0.00-3.00)	1.09 ± 0.74	1.00 (0.00-3.00)
HDL	44.1 ± 11.5	42.5 (23.0-90.4)	48.1 ± 11.0	46.1 (27.0-93.3)
LDL	117.1 ± 36.5	114.3 (22.0-217.0)	121.8 ± 38.1	119.6 (9.1-241.6)
Trigliserid	173.1 ± 94.7	152.0 (42.0-664.0)	130.7 ± 75.8	108.0 (34.0-426.0)
Kolesterol Yüksekliği	196.2 ± 42.5	194.0 (72.0-344.0)	196.6 ± 44.4	192.0 (101.0-367.0)
Açlık Kan Şekeri	191.8±99.2	161.0 (67.0-538.0)	95.4 ± 15.7	94.0 (69.0-199.0)

Tez verisinde bulunan bağımsız değişkenlerden sayısal olanların tanımlayıcı istatistik tablosu ayrıntılı olarak Tablo 4.4'de gösterilmiştir.

Tablo 4.4. Araştırmadaki nitel bağımsız değişkenlere ilişkin tanımlayıcı istatistiksel ölçütlerin dağılımı

		Tip 2 Diyabet Mellitus: Var n=146	**Tip 2 Diyabet Mellitus: Yok n=167**
Değişken	**Kategori**	**Sayı (%)**	**Sayı (%)**
Cinsiyet	Kadın	70 (47.9)	124 (74.3)
	Erkek	76 (52.1)	43 (25.7)
Aile Öyküsü	Var	80 (54.8)	72 (43.1)
	Yok	66 (45.2)	95 (56.9)
Uzun süre ilaç kullanımı	Var	37 (25.3)	37 (22.2)
	Yok	109 (74.7)	130 (77.8)
Kortizon kullanımı	Var	4 (2.7)	3 (1.8)
	Yok	142 (97.3)	164 (98.2)
Eşlik eden hastalık	Var	97 (66.4)	65 (38.9)
	Yok	49 (33.6)	102 (61.1)
Yüksek tansiyon	Var	65 (44.5)	35 (21.0)
	Yok	81 (55.5)	132 (79.0)
Stres faktörü	Yok	27 (18.5)	39 (23.4)
	Az var	31 (21.2)	34 (20.4)
	Çok var	88 (60.3)	94 (56.3)
Kalp hastalığı	Var	51 (13.2)	22 (13.2)
	Yok	95 (86.8)	145 (86.8)
Kolesterol yüksekliği	Var	54 (37.0)	18 (10.8)
	Yok	92 (63.0)	149 (89.2)
Sigara kullanımı	İçiyor	34 (23.3)	40 (24.0)
	İçmiyor	112 (76.7)	127 (76.0)
Alkol tüketimi	Var	4 (2.7)	2 (1.2)
	Yok	142 (97.3)	165 (98.8)
Egzersiz Durumu	Düzenli	23 (15.8)	20 (12.0)
	Düzensiz	98 (67.1)	128 (76.6)
	Arasıra	25 (17.1)	19 (11.4)
Karbonhidrat kullanımı	Var	94 (64.4)	89 (53.3)
	Yok	52 (35.6)	78 (46.7)
Sebze kullanımı	Var	124 (84.9)	141 (84.4)
	Yok	22 (15.1)	26 (15.6)
Et kullanımı	Var	95 (65.1)	105 (62.9)
	Yok	51 (34.9)	62 (37.1)

Kullandığımız üç modelden Yapay Sinir Ağı modeline parametre optimizasyon yöntemi uygulanmadan önceki ve sonraki sınıflandırma performansları belirlenen performans ölçütlerine göre, Tablo 4.5,Tablo 4.6, Tablo 4.7 ve Tablo 4.8'de verilmiştir.

Tablo 4.5. Lojistik Regresyon, Yapay Sinir Ağları ve Karar Ağaçları yöntemlerinden Yapay Sinir Ağı modeline parametre optimizasyon yöntemi uygulanmadan önceki eğitim verisi için sınıflandırma performansı

Yöntem	Doğruluk	Duyarlılık	Seçicilik	Kesinlik	F-ölçümü	AUC	Sınıflama Hatası
Lojistik Regresyon	90.04	94.78	84.62	87.59	91.04	0.954	9.96
Yapay Sinir Ağları	83.56	89.74	76.47	81.40	85.37	0.882	16.44
Karar Ağaçları	94.02	97.76	89.74	91.61	94.58	0.954	5.98

Tablo 4.5'e göre eğitim verisi için Lojistik Regresyon yönteminin sınıflandırma performans değerleri doğruluk, duyarlılık, seçicilik, kesinlik, F-ölçümü, AUC ve sınıflama hatası ölçütleri açısından sırasıyla 90.04, 94.78, 84.62, 87.59, 91.04, 0.954, 9.96; Yapay Sinir Ağları yönteminin sınıflandırma performans değerleri doğruluk, duyarlılık, seçicilik, kesinlik, F-ölçümü, AUC ve sınıflama hatası ölçütleri açısından sırasıyla 83.56, 89.74, 76.47, 81.40, 85.37, 0.882, 16.44; Karar Ağaçları yönteminin sınıflandırma performans değerleri doğruluk, duyarlılık, seçicilik, kesinlik, F-ölçümü, AUC ve sınıflama hatası ölçütleri açısından sırasıyla 94.02, 97.76, 89.74, 91.61, 94.58, 0.954, 5.98 olarak bulunmuştur.

Tablo 4.6. Lojistik Regresyon, Yapay Sinir Ağları ve Karar Ağaçları yöntemlerinden Yapay Sinir Ağı modeline parametre optimizasyon yöntemi uygulandıktan sonraki eğitim verisi için sınıflandırma performansı

Yöntem	Doğruluk	Duyarlılık	Seçicilik	Kesinlik	F-ölçümü	AUC	Sınıflama Hatası
Lojistik Regresyon	90.04	94.78	84.62	87.59	91.04	0.954	9.96
Yapay Sinir Ağları	98.63	100.00	97.06	97.50	98.73	0.977	1.37
Karar Ağaçları	94.02	97.76	89.74	91.61	94.58	0.954	5.98

Tablo 4.6'ya göre eğitim verisi için Lojistik Regresyon yönteminin sınıflandırma performans değerleri doğruluk, duyarlılık, seçicilik, kesinlik, F-ölçümü, AUC ve sınıflama hatası ölçütleri açısından sırasıyla 90.04, 94.78, 84.62, 87.59, 91.04, 0.954, 9.96; Yapay Sinir Ağları yönteminin sınıflandırma performans değerleri doğruluk, duyarlılık, seçicilik, kesinlik, F-ölçümü, AUC ve sınıflama hatası ölçütleri açısından sırasıyla 98.63, 100.00, 97.06, 97.50, 98.73, 0.977, 1.37; Karar Ağaçları yönteminin sınıflandırma performans değerleri doğruluk, duyarlılık, seçicilik, kesinlik, F-ölçümü, AUC ve sınıflama hatası ölçütleri açısından sırasıyla 94.02, 97.76, 89.74, 91.61, 94.58, 0.954, 5.98 olarak bulunmuştur.

Tablo 4.7. Lojistik Regresyon, Yapay Sinir Ağları ve Karar Ağaçları yöntemlerinden Yapay Sinir Ağı modeline parametre optimizasyon yöntemi uygulanmadan önceki test verisi için sınıflandırma performansı

Yöntem	Doğruluk	Duyarlılık	Seçicilik	Kesinlik	F-ölçümü	AUC	Sınıflama Hatası
Lojistik Regresyon	82.26	84.85	79.31	82.35	83.58	0.874	17.74
Yapay Sinir Ağları	75.53	78.00	72.73	76.47	77.23	0.855	24.47
Karar Ağaçları	85.48	90.91	79.31	83.33	86.96	0.878	14.52

Tablo 4.7'ye göre test verisi için Lojistik Regresyon yönteminin sınıflandırma performans değerleri doğruluk, duyarlılık, seçicilik, kesinlik, F-ölçümü, AUC ve sınıflama hatası ölçütleri açısından sırasıyla 82.26, 84.85, 79.31, 82.35, 83.58, 0.874, 17.74; Yapay Sinir Ağları yönteminin sınıflandırma performans değerleri doğruluk, duyarlılık, seçicilik, kesinlik, F-ölçümü, AUC ve sınıflama hatası ölçütleri açısından sırasıyla 75.53, 78.00, 72.73, 76.47, 77.23, 0.855, 24.47; Karar Ağaçları yönteminin sınıflandırma performans değerleri doğruluk, duyarlılık, seçicilik, kesinlik, F-ölçümü, AUC ve sınıflama hatası ölçütleri açısından sırasıyla 85.48, 90.91, 79.31, 83.33, 86.96, 0.878, 14.52 olarak bulunmuştur.

Tablo 4.8. Lojistik Regresyon, Yapay Sinir Ağları ve Karar Ağaçları yöntemlerinden Yapay Sinir Ağı modeline parametre optimizasyon yöntemi uygulandıktan sonraki test verisi için sınıflandırma performansı

Yöntem	Doğruluk	Duyarlılık	Seçicilik	Kesinlik	F-ölçümü	AUC	Sınıflama Hatası
Lojistik Regresyon	82.26	84.85	79.31	82.35	83.58	0.874	17.74
Yapay Sinir Ağları	98.94	100.00	97.73	98.04	99.01	0.978	1.06
Karar Ağaçları	85.48	90.91	79.31	83.33	86.96	0.878	14.52

Tablo 4.8'e göre test verisi için Lojistik Regresyon yönteminin sınıflandırma performans değerleri doğruluk, duyarlılık, seçicilik, kesinlik, F-ölçümü, AUC ve sınıflama hatası ölçütleri açısından sırasıyla 82.26, 84.85, 79.31, 82.35, 83.58, 0.874, 17.74; Yapay Sinir Ağları yönteminin sınıflandırma performans değerleri doğruluk, duyarlılık, seçicilik, kesinlik, F-ölçümü, AUC ve sınıflama hatası ölçütleri açısından sırasıyla 98.94, 100, 97.73,98.04, 99.01, 0.978, 1.06; Karar Ağaçları yönteminin sınıflandırma performans değerleri doğruluk, duyarlılık, seçicilik, kesinlik, F-ölçümü, AUC ve sınıflama hatası ölçütleri açısından sırasıyla 85.48, 90.91, 79.31, 83.33, 86.96, 0.878, 14.52 olarak bulunmuştur.

Lojistik Regresyon, Yapay Sinir Ağları ve Karar Ağaçları yöntemlerine göre hangi bağımsız değişkenin bağımlı değişken üzerinde daha etkili olduğunu belirlemek amacıyla kullanılabilen ağırlıklar Tablo 4.9'da verilmiştir.

Tablo 4.9. Tip 2 DM'deki bağımsız değişkenlerin ağırlıklarının dağılımı

Değişkenler	**Lojistik Regresyon**	**Yapay Sinir Ağları**	**Karar Ağaçları**
Cinsiyet	0.782	0.017	0.176
Aile Öyküsü	0.482	0.013	0.044
Uzun süre ilaç kullanımı	0.160	0.009	-
Kortizon Kullanımı	0.232	0.008	-
Eşlik eden hastalık	0.742	0.017	0.097
Yüksek tansiyon	0.330	0.008	-
Stres faktörü	0.031	0.016	-
Kalp hastalığı	0.003	0.024	-
Kolestrol yüksekliği	0.992	0.053	-
Sigara kullanımı	0.186	0.006	-
Alkol tüketimi	0.172	0.007	-
Egzersiz durumu	0.089	0.023	-
Karbonhidrat Kullanımı	0.709	0.040	-
Sebze Kullanımı	0.220	0.020	-
Et kullanımı	0.497	0.007	-
Yaş	0.287	0.046	0.117
Kilo	0.458	0.083	0.340
Boy	0.155	0.049	-
Başlama yaşı	0.974	0.024	0.018
Günlük ekmek tüketimi	0.752	0.066	0.065
Yüksek yoğunluklu lipoprotein (HDL)	2.395	0.083	-
Düşük yoğunluklu lipoprotein (LDL)	4.897	0.084	-
Trigliserid	2.618	0.031	-
Total Kolesterol	6.588	0.020	-
Açlık Kan Şekeri	5.108	0.244	0.297

Tablo 4.9'a göre Lojistik Regresyon yöntemi için cinsiyet, aile öyküsü, uzun süre ilaç kullanımı, kortizon kullanımı, eşlik eden hastalık, yüksek tansiyon, stres faktörü, kalp hastalığı, kolesterol yüksekliği, sigara kullanımı, alkol tüketimi, egzersiz durumu, karbonhidrat kullanımı, sebze kullanımı, et kullanımı, yaş, kilo, boy, başlama yaşı, günlük ekmek tüketimi, HDL, LDL, Trigliserid, Total Kolesterol, Açlık kan şekeri bağımsız değişkenlerinin ağırlıkları sırasıyla; 0.782, 0.482, 0.160, 0.232, 0.742, 0.330, 0.031, 0.003, 0.992, 0.186, 0.172, 0.089, 0.709, 0.220, 0.497, 0.287, 0.458, 0. 155, 0.974, 0.752, 2.395, 4.897, 2.618, 6.588, 5.108; yapay sinir ağları yöntemi için cinsiyet, aile öyküsü, uzun süre ilaç kullanımı, kortizon kullanımı, eşlik eden hastalık, yüksek tansiyon, stres faktörü, kalp hastalığı, kolesterol yüksekliği, sigara kullanımı, alkol tüketimi, egzersiz durumu, karbonhidrat kullanımı, sebze kullanımı, et kullanımı, yaş, kilo, boy, başlama yaşı, günlük ekmek tüketimi, HDL, LDL, Trigliserid, Total Kolesterol, Açlık kan şekeri bağımsız değişkenlerinin ağırlıkları sırasıyla; 0.017, 0.013, 0.009, 0.008, 0.017, 0.008, 0.016, 0.024, 0.053, 0.006, 0.007, 0.023, 0.040, 0.020, 0.007, 0.046, 0.083, 0.049, 0.024, 0.066, 0.083, 0.084, 0.031, 0.020, 0.244; karar ağaçları yöntemi için cinsiyet, aile öyküsü, eşlik eden hastalık, yaş, kilo, başlama yaşı, günlük ekmek tüketimi, açlık kan şekeri bağımsız değişkenlerinin ağırlıkları sırasıyla; 0.176, 0.044, 0.097, 0.117, 0.340, 0.018, 0.065, 0.297 olarak bulunmuştur.

5. TARTIŞMA

Çalışmada bağımlı değişken olarak Tip 2 DM tanısı almış ve almamış hastalar kullanılırken, bağımsız değişkenler ise cinsiyet, aile öyküsü, uzun süre ilaç kullanımı, kortizon kullanımı, eşlik eden hastalık, yüksek tansiyon, stres faktörü, kalp hastalığı, kolesterol yüksekliği, sigara kullanımı, alkol tüketimi, egzersiz durumu, karbonhidrat kullanımı, sebze kullanımı, et kullanımı, yaş, kilo, boy, başlama yaşı, günlük ekmek tüketimi, HDL, LDL, Trigliserid, Total Kolesterol, Açlık kan şekeri olarak belirlenmiştir. Sınıflandırma modellerinden Yapay Sinir Ağları, Lojistik Regresyon Analizi ve Karar Ağaçları yöntemlerinin performans karşılaştırılması bu veri seti üzerinde yapılmıştır. LRA modelinde kullanılan bağımsız değişkenler sınıflandırma performansı yönünden karşılaştırılacak olan YSA modelinde girdi olarak kabul edilir ve model buna göre oluşturulur. Karar ağaçları modeli için de bağımlı değişken Tip 2 DM olan ve olmayan hastalar olmak üzere belirlenen bağımsız değişkenler de kullanılarak ağaç oluşturulmuş ve analizler bu koşullar altında gerçekleştirilmiştir. LRA da bağımlı değişkenler ile bağımsız değişkenler arasında bulunan ilişki risk yönünden incelenebilir (44).YSA da ise beyindeki bir işlevi yerine getirebilmesi için kullanılan yöntemi modellemek için tasarlanmış olan matematiksel sistem yardımı ile oluşturulmuş modelin üzerinden sınıflandırılma yapılmaktadır.

Çalışmada kullanılan veri seti için yöntemlerin Tip 2 DM olan ve olmayan hastaları sınıflandırmada belirlenen performans ölçütleri (doğruluk, duyarlılık, seçicilik, kesinlik, F-ölçümü, AUC ve sınıflama hatası) dikkate alındığında, üç sınıflandırma modeline ait bulgular birlikte ele alındığında en iyi sınıflandırma performansı Yapay Sinir Ağları modeline ait olduğu belirlenmiştir. YSA modelinde; doğruluk 98.94, duyarlılık 100, seçicilik 97.73, kesinlik 98.04, F-ölçümü 99.01, AUC 0.978, sınıflama hatası 1.06 olarak elde edilmiştir.

Ayrıca, Tip 2 DM ile ilişkili faktörler, YSA, LRA, Karar Ağaçları ile belirlenerek, faktörlerin hastalık üzerindeki etkileri tahmin edilmiştir. En iyi performans sonucunu veren YSA modeline göre Tip 2 DM neden olabilecek en önemli faktör açlık kan şekeri olarak elde edilmiştir. Klinik olarak, artmış açlık kan şekerinin yani 126 mg/dl veya daha yüksek olan açlık kan değerinin Tip 2 DM üzerinde etkisi vardır (36).

Tip 2 DM sınıflandırılmasında, bu hastalığı etkileyebilecek en önemli ikinci faktörün LDL değeri olduğu belirlenmişken yine kan lipidlerinden olan HDL değeri de üçüncü sıradaki önemli risk faktörü olarak belirlenmiştir. Kolesterol ve Koroner kalp hastalıklarına ait ilişkilerde temel olarak etkili olan faktör LDL kolesterol değeridir ve Tip 2 diyabet mellitus

da risk açısından önemli bir belirleyicidir. Ek olarak HDL kolesterol düzeyi düşük olan kişiler Tip 2 diyabet mellitus için riski altındadırlar (36, 45).

Diğer önemli bir faktörün de kilo değişkeni olduğu belirlenmiştir. Beden Kütle İndeksi (BKİ) 25 kg/m^2'nin yukarısında olanlar ve özellikle göbek çevresi geniş olanları Tip 2 diyabet mellitus açısından risk taşımaktadırlar (36).

6. SONUÇ VE ÖNERİLER

Çalışmada yer alan Tip 2 DM olan ve olmayan hastalara ait veriler ve hastalığa etkisi olabilecek risk faktörlerinden oluşan veri seti için yapılan sınıflandırma performansını karşılaştırmaya yönelik analizlerde sınıflandırma performans ölçütlerinden olan doğruluk, duyarlılık, seçicilik, kesinlik, F-ölçümü, AUC ve sınıflama hatası için en iyi sonucun üç modelden YSA modelinin verdiği gözlemlenmiştir.

Ayrıca en iyi sınıflandırma performansını veren YSA modeline göre Tip 2 DM için risk faktörleri de bu yöntem ile belirlenmiş olup, hastalığa sebep olan en önemli risk faktörünün açlık kan şekeri değeri olduğu elde edilmiştir. Diğer bazı önemli risk faktörleri ise sırasıyla LDL, HDL, kilo şeklinde olup Tip 2 DM hastalığı için daha önce yapılan çalışmalarla paralel sonuçlar bulunmuştur.

Günümüzde yapay sinir ağları modeli çoğu alanda uygulandığı ve sınıflandırma modellerinde de başarılı sonuçların elde edildiği görülmektedir. YSA kullanımı esnek, hızlı, kolay ve tutarlı sonuçlar verdiği için tercih edilmiştir.

Bu konu ile ilgili yapılacak olan çalışmalarda; mevcut çalışmada ele alınan veri setinin yapısına uygun olan, farklı makine öğrenmesi algoritmaları ve veri madenciliği alanlarında sıklıkla kullanılan diğer yöntemlerle sınıflandırma performansı incelenerek tahmin başarısının daha da arttırılabilmesi önerilmektedir.

KAYNAKLAR

1. Akpınar H. Veri tabanlarında bilgi keşfi ve veri madenciliği. *İÜ İşletme Fakültesi Dergisi* 2000; 29(1): 1-22.

2. Köktürk F, Ankarali H, Sümbüloğlu V. Veri Madenciliği Yöntemlerine Genel Bakış. *Turkiye Klinikleri Journal of Biostatistics* 2009; 1(1): 20-5.

3. Güneri N, Apaydın A. Öğrencİ Başarilarinin Siniflandirilmasinda Lojİstİk Regresyon Analİzİ ve Sİnİr AğlarI Yaklaşimi. *Ticaret ve Turizm Eğitim Fakültesi Dergisi* 2004; 1170-88.

4. Kaya E, Bulun M, Arslan A. Tıpta Veri Ambarları Oluşturma ve Veri Madenciliği Uygulamaları. *Selçuk Üniversitesi, Bilgisayar Mühendisliği Bölümü* 2003.

5. Berry MJ, Linoff G. *Data mining techniques: for marketing, sales, and customer support*, John Wiley & Sons, Inc 1997.

6. Holsheimer M, Siebes A. *Data mining: The search for knowledge in databases*, Centrum voor Wiskunde en Informatica 1994.

7. Türe M. Ömürlü K..". *Sınıflandırma Yöntemlerinin Performanslarının Karşılaştırılmasına İlişkin Simülasyon Çalışması* 2009.

8. Fayyad U, Piatetsky-Shapiro G, Smyth P. From data mining to knowledge discovery in databases. *AI magazine* 1996; 17(3): 37.

9. Kurt İ, Türe M. Tıp Öğrencilerinde Alkol Kullanımını Etkileyen Faktörlerin Belirlenmesinde Yapay Sinir Ağları ile Lojistik Regresyon Analizi'nin Karşılaştırılması. *Trakya Üniversitesi Tıp Fak Dergisi* 2005; 22(3): 142-53.

10. Kıran Z. Lojistik Regresyon ve Cart Analizi Teknikleriyle Sosyal Güvenlik Kurumu Ġlaç Provizyon Sistemi Verileri Üzerinde Bir Uygulama Fen Bilimleri Enstitüsü, İstatistik Anabilim Dalı. Yükseklisans Tezi, Ankara: Gazi Üniversitesi 2010.

11. Karakış R. Yapay Sinir Ağları ve Lojistik Regresyon Yöntemleri ile Meme Kanseri Koltuk Altı Lenf Durumunun Belirlenmesi. Yüksek Lisans Tezi, Ankara, Gazi Üniversitesi 2009.

12. Alpar R. *Spor, sağlık ve eğitim bilimlerinden örneklerle uygulamalı istatistik ve geçerlik-güvenirlik*, Detay Yayıncılık 2010.

13. Hand DJ. Data mining: Statistics and more? *The American Statistician* 1998; 52(2): 112-8.

14. Zurada J, Lonial S. Comparison of the performance of several data mining methods for bad debt recovery in the healthcare industry. *Journal of Applied Business Research* 2005; 21(2): 37-54.

15. Han J, Kamber M, Pei J. *Data mining: concepts and techniques*. Morgan Kaufmann 2006: 230-40.

16. Tatlıdil H. *Uygulamalı çok değişkenli istatistiksel analiz*, Engin Yayınları 1992.

17. Hosmer Jr DW, Lemeshow S, Sturdivant RX. *Applied logistic regression*, John Wiley & Sons 2013.

18. Kuyucu YE. Lojistik regresyon analizi (LRA), yapay sinir ağları (YSA) ve sınıflandırma ve regresyon ağaçları (C&RT) yöntemlerinin karşılaştırılması ve tıp alanında bir uygulama. Sağlık Bilimleri Enstitüsü,Biyoistatistik Anabilim Dalı. Yüksek Lisans Tezi, Tokat: Gaziosmanpaşa Üniversitesi 2012.

19. Şahin ŞÖ. Yapay sinir ağları yardımı ile dinamik bir senaryo analizi. Fen Bilimleri Enstitüsü, . Doktora Tezi, İstanbul: İstanbul Teknik Üniversitesi 2001.

20. Jacobs P. Data mining: What general managers need to know. *Harvard Management Update* 1999; 4(10): 8.

21. Ocakoğlu G. Lojistik Regresyon Analizi ve Yapay Sinir Ağları Yöntemlerinin Sınıflama Özelliklerini Karşılaştırılması ve Bir Uygulama. Sağlik Bilimleri Enstitüsü, Biyoistatistik Anabilim Dalı Yüksek Lisans Tezi, Bursa: Uludağ Üniversitesi 2006.

22. Kaynak O, Efe M. *Yapay Sinir Ağları ve Uygulamaları*. Boğaziçi Üniversitesi 2000.

23. Fausett LV. *Fundamentals of neural networks: architectures, algorithms, and applications*, Prentice-Hall Englewood Cliffs 1994.

24. Seven A. Yapay Sinir Ağları ile Doku Sınıflandırma. Fen Bilimleri Enstitüsü, Elektrik ve Elektronik Mühendisliği. Yüksek Lisans Tezi, İstanbul: İstanbul Teknik Üniversitesi 1993.

25. Çolak MC, Çolak C, Kocatürk H, Sagiroglu S, Barutçu İ. Predicting coronary artery disease using different artificial neural network models/Koroner arter hastaliginin degisik yapay sinir agi modelleri ile tahmini. *Anadulu Kardiyoloji Dergisi: AKD* 2008; 8(4): 249.

26. Tosun S. Sınıflandırmada Yapay Sinir Ağları ve Karar Ağaçları Karşılaştırması: Öğrenci Başarıları Üzerine Bir Uygulama. Yüksek Lisans Tezi, İstanbul: İstanbul Teknik Üniversitesi 2007.

27. Öztemel E. *Yapay Sinir Ağlari*. Istanbul, PapatyaYayincilik 2012.

28. Aydoğan F. E-Ticarette Veri Madenciliği Yaklaşımlarıyla Müşteriye Hizmet Sunan Akıllı Modüllerin Tasarımı Ve Gerçekleştirimi. Fen Bilimleri Enstitüsü, Bilgisayar Mühendisliği. Yüksek Lisans Tezi, Ankara: Hacettepe Üniversitesi 2003.

29. Altıntaş Y. Veri Madenciliğinin Tıpta Kullanımı Ve Bir Uygulama: Hemodiyaliz Hastaları İçin Risk Seviyelerine Göre Risk Faktörlerinin Etkileşimlerinin İncelemesi. Fen Bilimleri Enstitüsü, Endüstri Mühendisliği. Yüksek Lisans Tezi, Ankara: Gazi Üniversitesi 2010.

30. Pehlivan G. Chaid Analizi ve Bir Uygulama. Fen Bilimleri Enstitüsü, İstatistik Anabilim Dalı. Yüksek LisansTezi, İstanbul: Yıldız Teknik Üniversitesi 2006.

31. Thomas LC. A survey of credit and behavioural scoring: forecasting financial risk of lending to consumers. *INT J FORECASTING* 2000; 16(2): 149-72.

32. Silahtaroğlu G. *Veri madenciliği*. Papatya Yayınları 2013.

33. Temel GO, Çamdeviren H, Akkuş Z. Sınıflama ağaçları yardımıyla Restless legs syndrome (RLS) hastalarına tanı koyma. *Turgut Özal Tıp Merkezi Dergisi* 2005; 12(2): 111-7.

34. Vahaplar A. Bir Coğrafi Veri Madenciliği Uygulaması. Fen Bilimler Enstitüsü, Bilgisayar Mühendisliği Anabilim Dalı. Yüksek Lisans Tezi, İzmir: Ege Üniversitesi 2003.

35. Masseglia F, Poncelet P, Teisseire M. Using data mining techniques on web access logs to dynamically improve hypertext structure. *ACM sigweb Newsletter* 1999; 8(3): 13-9.

36. Yılmaz T. Diabetes Mellitusun Tanı Kriterleri Ve Sınıflaması. *Türkiye Diyabet Vakfı* 2003; 1.

37. Rapidminer DA. RapidMiner 4.1 User Guide. Dortmund; 2008.

38. de Andrade Silva J, Hruschka ER. An experimental study on the use of nearest neighbor-based imputation algorithms for classification tasks. *Data & Knowledge Engineering* 2013; 8447-58.

39. Yılmaz H. Random Forests Yönteminde Kayıp Veri Probleminin Incelenmesi Ve Sağlık Alanında Bir Uygulama. Sağlık Bilimleri Enstitüsü, Biyoistatistik Anabilim Dalı. Yüksek Lisans Tezi, Eskişehir: Eskişehir Osmangazi Üniversitesi 2014.

40. Land S, Fischer S. RapidMiner 5: RapidMiner in academic use. *Rapid-I GmbH, Dortmund, Germany* 2012.

41. Akthar F, Hahne C. *Rapidminer 5 operator reference* 2012: 65.

42. Coşkun C, Baykal A. Veri Madenciliğinde Sınıflandırma Algoritmalarının Bir Örnek Üzerinde Karşılaştırılması. *Akademik Bilişim* 2011; 1(1): 1-8.

43. Dirican A. Tanı testi performanslarının değerlendirilmesi ve kıyaslanması. *Cerrahpaşa J Med* 2001; 32(1): 25-30.
44. Ediz B. Lojistik Regresyon-Ayırma Analizi, Ayrımsama Sorunu ve Kalp Hastalarında Lojistik Model Yardımıyla Risk Ölçütlerinin Belirlenmesi. Sağlık Bilimleri Enstitüsü. Doktora Tezi, Bursa: Uludağ Üniversitesi 1997.
45. Özdoğan E, Özdoğan O, Altunoğlu EG, Köksal AR. Tip 2 diyabet hastalarında kan lipid düzeylerinin Hba1c ve obezite ile ilişkisi. *Şişli Etfal Tıp Bülteni* 2007; 49(4): 248-54.

EKLER

EK-1. Özgeçmiş

ÖZGEÇMİŞ

Adı Soyadı: İpek BALIKÇI ÇİÇEK

Doğum Tarihi: 1988

Öğrenim Durumu: Yüksek Lisans

Derece	Bölüm/Program	Üniversite	Yıl
Lisans	Matematik	Çukurova Üniversitesi	2010
Y. Lisans	Biyoistatistik ve Tıp Bilişimi AD	İnönü Üniversitesi	2018

Yüksek Lisans Tez Başlığı ve Tez Danışman(lar)ı:

Tip 2 Diyabet Mellitus İle İlişkili Risk Faktörlerini Saptamada Çok Değişkenli İstatistiksel Yöntemlerinin Karşılaştırılması, Prof. Dr. Saim YOLOĞLU

Görevler:

Görev Unvanı	Görev Yeri	Yıl
Arş. Gör.	İnönü Üniversitesi	2014 – Devam ediyor

EK-2. Etik Kurul Onay Formu

KLİNİK ARAŞTIRMALAR ETİK KURULU KARAR FORMU

ARAŞTIRMANIN AÇIK ADI	Tip 2 Diyabet Mellitus İle İlişkili Risk Faktörlerini Saptamada Çok Değişkenli İstatistiksel Yöntemlerinin Karşılaştırılması
VARSA ARAŞTIRMANIN PROTOKOL KODU	2016/144

ETİK KURUL BİLGİLERİ	ETİK KURULUN ADI	MALATYA KLİNİK ARAŞTIRMALAR ETİK KURULU
	AÇIK ADRESİ:	İnönü Üniversitesi Merkez Kampüsü, 44280, Malatya, Türkiye
	TELEFON	+90 422 341 06 60 / 1219
	FAKS	+90 422 341 00 36
	E-POSTA	inu.dhek@inonu.edu.tr

BAŞVURU BİLGİLERİ	KOORDİNATÖR/SORUMLU ARAŞTIRMACI UNVANI/ADI/SOYADI	Prof. Dr. Saim Yoloğlu			
	KOORDİNATÖR/SORUMLU ARAŞTIRMACININ UZMANLIK ALANI	İnönü Üniversitesi Tıp Fakültesi Biyoistatistik ve Tıp Bilişimi AD			
	KOORDİNATÖR/SORUMLU ARAŞTIRMACININ BULUNDUĞU MERKEZ	MALATYA			
	VARSA İDARİ SORUMLU UNVANI/ADI/SOYADI				
	DESTEKLEYİCİ				
	PROJE YÜRÜTÜCÜSÜ UNVANI/ADI/SOYADI (TÜBİTAK vb. gibi kaynaklardan destek alanlar için)				
	DESTEKLEYİCİNİN YASAL TEMSİLCİSİ				
	ARAŞTIRMANIN FAZI VE TÜRÜ	FAZ 1	☐		
		FAZ 2	☐		
		FAZ 3	☐		
		FAZ 4	☐		
		Gözlemsel ilaç çalışması	☐		
		Tıbbi cihaz klinik araştırması	☐		
		İn vitro tıbbi tanı cihazları ile yapılan performans değerlendirme çalışmaları	☐		
		İlaç dışı klinik araştırma	☐		
		Diğer ise belirtiniz			
	ARAŞTIRMAYA KATILAN MERKEZLER	TEK MERKEZ ☐	ÇOK MERKEZLİ ☐	ULUSAL ☐	ULUSLARARASI ☐

Etik Kurul Başkanının
Unvanı/Adı/Soyadı: Prof. Dr. Rıfat KARLIDAĞ
İmza:

Not: Etik kurul başkanının her sayfada imzasının olması gerekmektedir.

KLİNİK ARAŞTIRMALAR ETİK KURULU KARAR FORMU

ARAŞTIRMANIN AÇIK ADI	Tip 2 Diyabet Mellitus İle İlişkili Risk Faktörlerini Saptamada Çok Değişkenli İstatistiksel Yöntemlerinin Karşılaştırılması
VARSA ARAŞTIRMANIN PROTOKOL KODU	2016/144

	Belge Adı	Tarihi	Versiyon Numarası	Dili
DEĞERLENDİRİLEN BELGELER	ARAŞTIRMA PROTOKOLÜ			Türkçe ☐ İngilizce ☐ Diğer ☐
	BİLGİLENDİRİLMİŞ GÖNÜLLÜ OLUR FORMU			Türkçe ☐ İngilizce ☐ Diğer ☐
	OLGU RAPOR FORMU			Türkçe ☐ İngilizce ☐ Diğer ☐
	ARAŞTIRMA BROŞÜRÜ			Türkçe ☐ İngilizce ☐ Diğer ☐

	Belge Adı		Açıklama
DEĞERLENDİRİLEN DİĞER BELGELER	SİGORTA	☐	
	ARAŞTIRMA BÜTÇESİ	☐	
	BİYOLOJİK MATERYEL TRANSFER FORMU	☐	
	İLAN	☐	
	YILLIK BİLDİRİM	☐	
	SONUÇ RAPORU	☐	
	GÜVENLİLİK BİLDİRİMLERİ	☐	
	DİĞER:	☐	
KARAR BİLGİLERİ	**Karar No:2016/144**	**Tarih:22.06.2016**	
	Yukarıda bilgileri verilen başvuru dosyası ile ilgili belgeler araştırmanın/çalışmanın gerekçe, amaç, yaklaşım ve yöntemleri dikkate alınarak incelenmiş ve uygun bulunmuş olup araştırmanın/çalışmanın başvuru dosyasında belirtilen merkezlerde gerçekleştirilmesinde etik ve bilimsel sakınca bulunmadığına toplantıya katılan etik kurul üye tam sayısının salt çoğunluğu ile karar verilmiştir. İlaç ve Biyolojik Ürünlerin Klinik Araştırmaları Hakkında Yönetmelik kapsamında yer alan araştırmalar/çalışmalar için Türkiye İlaç ve Tıbbi Cihaz Kurumu'ndan izin alınması gerekmektedir.		

KLİNİK ARAŞTIRMALAR ETİK KURULU	
ETİK KURULUN ÇALIŞMA ESASI	İlaç ve Biyolojik Ürünlerin Klinik Araştırmaları Hakkında Yönetmelik, İyi Klinik Uygulamaları Kılavuzu
BAŞKANIN UNVANI / ADI / SOYADI:	Prof. Dr. Rıfat KARLIDAĞ

Unvanı/Adı/Soyadı	Uzmanlık Alanı	Kurumu	Cinsiyet		Araştırma ile ilişki		Katılım *		İmza
Prof. Dr. Rıfat KARLIDAĞ	Psikiyatri	İnönü Üniversitesi Tıp Fakültesi	E ☒	K ☐	E ☐	H ☒	E ☒	H ☐	
Prof. Dr. Metin GENÇ	Halk Sağlığı	İnönü Üniversitesi Tıp Fakültesi	E ☒	K ☐	E ☐	H ☒	E ☒	H ☐	
Prof. Dr. Saim YOLOĞLU	Biyoistatistik	İnönü Üniversitesi Tıp Fakültesi	E ☒	K ☐	E ☐	H ☐	E ☐	H ☐	Çekimser
Prof. Dr. Türkan TOĞAL	Anestezyoloji ve Rea.	İnönü Üniversitesi Tıp Fakültesi	E ☐	K ☒	E ☐	H ☒	E ☒	H ☐	
Prof. Dr. İbrahim ŞAHİN	İç Hastalıkları	İnönü Üniversitesi Tıp Fakültesi	E ☒	K ☐	E ☐	H ☒	E ☐	H ☐	Araştırmacı
Prof. Dr. Sedat YILDIZ	Fizyoloji	İnönü Üniversitesi Tıp Fakültesi	E ☒	K ☐	E ☐	H ☒	E ☒	H ☐	
Doç. Dr. Seda TAŞDEMİR	Tıbbi Farmakoloji	İnönü Üniversitesi Tıp Fakültesi	E ☐	K ☒	E ☐	H ☒	E ☒	H ☐	

Etik Kurul Başkanının
Unvanı/Adı/Soyadı: Prof. Dr. Rıfat KARLIDAĞ
İmza:

Not: Etik kurul başkanının her sayfada imzasının olması gerekmektedir.

KLİNİK ARAŞTIRMALAR ETİK KURULU KARAR FORMU

AŞTIRMANIN AÇIK ADI	Tip 2 Diyabet Mellitus İle İlişkili Risk Faktörlerini Saptamada Çok Değişkenli İstatistiksel Yöntemlerinin Karşılaştırılması
RSA ARAŞTIRMANIN OTOKOL KODU	2016/144

Doç. Dr. Derya DOĞAN	Çocuk Sağlığı ve Hast.	İnönü Üniversitesi Tıp Fakültesi	E☐	K☑	E☐	H☐	E☐	H☐	Katılmak
Doç. Dr. Özden KAMIŞLI	Nöroloji	İnönü Üniversitesi Tıp Fakültesi	E☐	K☑	E☐	H☑	E☑	H☐	
Doç. Dr. Hakan HARPUTLUOĞLU	Onkoloji	İnönü Üniversitesi Tıp Fakültesi	E☑	K☐	E☐	H☑	E☐	H☐	
Yrd. Doç. Dr. Mehmet KARATAŞ	Tıp Tarihi ve Etik	İnönü Üniversitesi Tıp Fakültesi	E☑	K☐	E☐	H☐	E☐	H☐	
Dr. Mahmut Barkın AKGÜL	Tıp Doktoru	Halk Sağlığı Müdürlüğü	E☑	K☐	E☐	H☐	E☑	H☐	
Metin TAY	Eczacı	Serbest Eczacı	E☑	K☐	E☐	H☐	E☐	H☐	Katılmak
Zafer ERGÜZEL	Hukuk	İnönü Üniversitesi	E☑	K☐	E☐	H☐	E☐	H☐	Katılmak
Hasan KONAN	Sivil Üye	MSD Ltd. Şti.	E☑	K☐	E☐	H☑	E☑	H☐	

Etik Kurul Başkanının
Unvanı/Adı/Soyadı: Prof. Dr. Rıfat KARLIDAĞ
İmza:

Not: Etik kurul başkanının her sayfada imzasının olması gerekmektedir.

EK-3. Anket Formu

DİYABET (TİP 2) İÇİN RİSK FAKTÖRLERİ

1. Yaşınız (Yıl)? .....
2. Cinsiyetiniz?
   1) Kadın 2) Erkek
3. Kilonuz .....kg
4. Boyunuz .....cm
5. Medeni Durumunuz?
   1) Evli 2) Bekar 3) Diğer
6. Gebelik sayınız? (Bayan hastalar için) .....
7. Doğum kilonuz? (Bayan hastalar için) .....
8. Gebelikte diyabet hastalığınız oldu mu? (Bayan hastalar için)
   1) Evet 2) Hayır
9. Eğitim Durumunuz nedir?
   1) Okur-yazar değil 2) Ortaokul mezunu 3) Okuryazar
   4) Lise mezunu 5) İlkokul mezunu 6) Üniversite mezunu
10. Mesleğiniz nedir?
    1) İşçi 2) Ev hanımı 3) Emekli
    4) Memur 5) Serbest meslek 6) İşsiz 7) Diğer
11. Yerleşim Yeriniz?
    1) Kırsal 2) Kentsel
12. Aylık evinize giren gelir ne kadar? .....
13. Diyabet hastalığı başlama yaşınız? .....
14. Ailede (anne, baba ve kardeşler…) diyabet hastası olan var mı?
    1) Evet 2) Hayır
15. Diyabet olmadan önce kullandığınız uzun süreli ilaçlar var mı?
    1)Var / Var ise isimleri nelerdir?.....
    2) Yok
16. Diyabet olmadan önce Kortizon kullandınız mı?
    1) Evet 2) Hayır
17. Diyabete eşlik eden başka bir hastalığınız var mı?
    1) Var 2) Yok
18. Hipertansiyonunuz var mı?
    1)Var/Var ise *Sistolik……..(mmHg) *Diyastolik…….(mmHg)
    2) Yok
19. Stresiniz var mı?
    1) Yok 2) Az var 3) Çok var
20. Kalp hastalığınız var mı?
    1) Var 2) Yok
21. Kolesterol yüksekliğiniz var mı?
    1) Var 2) Yok
22. Sigara içiyor musunuz?
    1) İçiyor / İçiyor iseniz kaç yıldır içiyorsunuz..... Günde kaç adet.....
    2) İçmiyor

23. Alkol tüketiminiz?
1) Var 2) Yok

24. Egzersiz Durumunuz?
1) Düzenli egzersiz yapıyor 2) Düzenli egzersiz yapmıyor 3) Arasıra yapıyor

25. Beslenme Şekliniz?
1) Karbonhidrat ağırlıklı 1) Evet 2) Hayır
2) Sebze Ağırlıklı 1) Evet 2) Hayır
3) Et Ağırlıklı 1) Evet 2) Hayır

26. Günlük ekmek miktarınız? .....

27. Adet düzensizliğiniz var mı? (Bayan hastalar için)
1) Var 2) Yok

28. Polikstik Over hastalığınız var mı? (Bayan hastalar için)
1) Var 2) Yok

29. HDL değeriniz.....

30. LDL değeriniz.....

31. Trigliserid değeriniz.....

32. Total Kolesterol değeriniz.....

33. Açlık kan şekeri ölçümünüz.....

34. Tokluk kan şekeri ölçümünüz.....

35. HbA1C ölçümünüz.....