

**T.C.**  
**AKDENİZ ÜNİVERSİTESİ**  
**SAĞLIK BİLİMLERİ ENSTİTÜSÜ**  
**BİYOİSTATİSTİK VE TIBBİ BİLİŞİM ANABİLİM DALI**

**Ayça ŞANLI**

**TİP 2 DİYABETİN ERKEN TANISINDA MAKİNE ÖĞRENME**  
**TEKNİKLERİNİN KULLANILMASI**

**Ayça ŞANLI**

**YÜKSEK LİSANS TEZİ**

**YÜKSEK LİSANS TEZİ**

**2023-ANTALYA**

**2023-ANTALYA**

**T.C.**  
**AKDENİZ ÜNİVERSİTESİ**  
**SAĞLIK BİLİMLERİ ENSTİTÜSÜ**  
**BİYOİSTATİSTİK VE TIBBİ BİLİŞİM ANABİLİM DALI**

**TİP 2 DİYABETİN ERKEN TANISINDA MAKİNE ÖĞRENME**  
**TEKNİKLERİNİN KULLANILMASI**

**Ayça ŞANLI**  
**YÜKSEK LİSANS TEZİ**

**DANIŞMAN**  
**Doç. Dr. Uğur BİLGE**

“Kaynakça gösterilerek tezimden yararlanılabilir”

2023-ANTALYA

**Sağlık Bilimleri Enstitüsü Müdürlüğüne;**

Bu çalışma jürimiz tarafından Biyoistatistik ve Tıbbi Bilişim Anabilim Dalı Tıp Bilişimi yüksek lisans Programında yüksek lisans tezi olarak kabul edilmiştir. .../...../.....

İmza

Tez Danışmanı : Doç. Dr. Uğur BİLGE  
Akdeniz Üniversitesi

.....

Üye : Doç. Dr. K. Hakan GÜLKESEN  
Akdeniz Üniversitesi

.....

Üye : Dr. Öğr. Üyesi Yılmaz Kemal YÜCE  
Alanya Alaaddin Keykubat Üniversitesi

.....

Bu tez, Enstitü Yönetim Kurulunca belirlenen yukarıdaki jüri üyeleri tarafından uygun görülmüş ve Enstitü Yönetim Kurulu'nun ...../...../..... tarih ve ...../..... sayılı kararıyla kabul edilmiştir.

Enstitü Müdürü

## ETİK BEYAN

Bu tez çalışmasının kendi çalışmam olduğunu, tezin planlanmasından yazımına kadar bütün safhalarda etik dışı davranışımın olmadığını, bu tezdeki bütün bilgileri akademik ve etik kurallar içinde elde ettiğimi, bu tez çalışmasıyla elde edilmeyen bütün bilgi ve yorumlara kaynak gösterdiğimi ve bu kaynakları da kaynaklar listesine aldığımı beyan ederim.

Öğrenci

Ayça ŞANLI

İmza

Tez Danışmanı

Doç. Dr. Uğur BİLGE

İmza

## TEŞEKKÜR

Bu tez çalışması boyunca çalışmama liderlik eden ve bana her zaman destek olan tez danışmanım sayın Doç Dr Uğur Bilge'ye teşekkürü bir borç bilirim. Üzerimde çok emeği olan, sevgiyi ve saygıyı öğreten ananem Şerife Çelikpençe'ye, dedem Kaya Çelikpençe'ye ve aileme teşekkür ederim.



## ÖZET

**Amaç:** Kronik bir metabolizma bozukluğu olan diyabet, pankreasın artık insülin üretemediği veya vücudun ürettiği insülini yeterince kullanamadığı durumlarda ortaya çıkan ve ömür boyu süren bir hastalıktır. Hastalığın ilerleyen aşamalarında göz, böbrek, kalp ve damar rahatsızlıklarını da meydana getiren diyabet, komplikasyonları önlemek veya geciktirmek, hastaların yaşam kalitesini iyileştirmekle beraber diyabetli olduklarının farkında olmayan diyabetli kişileri teşhis etmek ve diyabetli tüm insanlara mümkün olan en erken zamanda bakım sağlamak oldukça önemlidir. Bu çalışmada yapay zekâ teknolojisinden yararlanılarak tip 2 diyabet hastalığının önceden öngörülmesi konusunun ele alınması ve böylelikle erken tanının sağlık sistemlerindeki ek iş yükünü hafifletmesi, hastaların yaşam kalitesini iyileştirmesi ve yüksek tedavi giderleri açısından ekonomik olarak sıkıntıya düşmelerinin engellenmesi amaçlanmıştır.

**Yöntem:** Akdeniz, Marmara ve Ege coğrafi bölgelerindeki özel hastanelerden toplanan 126 hastaya (66 diyabetik olmayan vaka ve 60 diyabetik vaka) ait veriler incelenmiştir. Uzman hekimler tarafından Tip 2 Diyabet tanısı konulmuş hastalar ve hiç diyabet tanısı konulmamış kontrol hastaları olarak gruplanmıştır. Lojistik Regresyon (LR), Rastgele Orman (RO) ve Karar Ağaçları grubun tahmini için model oluşturmak için kullanılmıştır. Makine öğrenmesi algoritmalarının performanslarını karşılaştırmak için AUC değeri, doğruluk, duyarlılık, özgüllük, kesinlik ve F1 skoru kullanıldı.

**Bulgular:** Tip 2 diyabete sahip bireylere ve kontrol grubuna ait tam kan sayımı (hemogram) testi sonuçları ve sosyo-demografik özellikler kullanılmıştır. Lojistik Regresyon, Rastgele orman ve Karar ağacı teknikleri sırasıyla %92.6 (88.1-97.1), %73.6 (65.9-81.4) ve %84.5 (78.3-90.7) ROC AUC vermektedir. Lojistik Regresyon algoritması diğer makine öğrenmesi algoritmalarından daha yüksek bir başarı ile %92.6 ROC AUC ve 88.1-97.1 güven aralığı ile sınıflandırmayı gerçekleştirmiştir.

**Sonuç:** Gerçekleştirilen çalışma sonucunda yeni bir yöntem önerisi getirilmemiş olsa da bundan sonraki benzer çalışmalara bir temel oluşturmuş ve üzerinde çalışılan makine öğrenmesi algoritması için karşılaştırma yapabilmemizi sağlayacak veriler sunmaktadır. Tip 2 diyabet hastalığının tahmininde en başarılı yöntemin Lojistik Regresyon olduğu görülmüştür.

**Anahtar Kelimeler:** Tip 2 diabetes mellitus, tam kan sayımı, makine öğrenmesi, yapay zeka

## ABSTRACT

**Objective:** Diabetes, a chronic metabolic disorder, is a lifelong disease that occurs when the pancreas no longer produces enough insulin or when the body cannot effectively use the insulin it produces. In this study, the use of artificial intelligence technology to predict type 2 diabetes is addressed. The aim is to lighten the additional workload on healthcare systems due to early diagnosis, improve the quality of life of patients, and prevent them from experiencing economic difficulties due to high treatment costs.

**Method:** We examined data from 126 patients obtained from private hospitals in the Mediterranean, Marmara, and Aegean geographical regions. We grouped patients diagnosed with Type 2 Diabetes by specialist physicians and control patients without a diabetes diagnosis. Logistic Regression (LR), Random Forest (RO), and Decision Trees were used to create models for predicting the group. AUC value, accuracy, sensitivity, specificity, precision, and F1 score were used to compare the performances of machine learning algorithms.

**Results:** Complete blood count (hemogram) test results and sociodemographic characteristics were used for individuals with Type 2 diabetes and the control group. Logistic Regression, Random Forest, and Decision Tree techniques yield ROC AUC values of 92.6% (88.1-97.1), 73.6% (65.9-81.4), and 84.5% (78.3-90.7), respectively. The Logistic Regression algorithm achieved higher success in classification with a ROC AUC of 92.6% and 88.1-97.1 confidence interval compared to other machine learning algorithms.

**Conclusion:** Although no new method proposal was made in this study, it provides a foundation for future similar studies and presents data for comparing the machine learning algorithm studied. The most successful method for predicting Type 2 diabetes was found to be Logistic Regression.

**Keywords:** Type 2 diabetes mellitus, complete blood count, machine learning, artificial intelligence

## İÇİNDEKİLER

|                                                                             |     |
|-----------------------------------------------------------------------------|-----|
| <b>ÖZET</b> .....                                                           | i   |
| <b>ABSTRACT</b> .....                                                       | ii  |
| <b>İÇİNDEKİLER</b> .....                                                    | iii |
| <b>TABLolar DİZİNİ</b> .....                                                | v   |
| <b>ŞEKİLLER DİZİNİ</b> .....                                                | vi  |
| <b>SİMGELER ve KISALTMALAR</b> .....                                        | vii |
| <br>                                                                        |     |
| <b>1. GİRİŞ</b> .....                                                       | 6   |
| <b>2. GENEL BİLGİLER</b> .....                                              | 7   |
| 2.1. Diabetes Mellitus.....                                                 | 7   |
| 2.1.1. Diyabetin Tarihçesi .....                                            | 7   |
| 2.1.2. Diyabetin Tipleri .....                                              | 7   |
| 2.1.3. Diabetes mellitus için tanı testleri .....                           | 8   |
| 2.2. Yapay Zeka ve Makine Öğrenmesi .....                                   | 11  |
| 2.2.1. Makine Öğrenmesi .....                                               | 13  |
| 2.2.2. Makine Öğrenimi Görevleri ve Algoritmaları.....                      | 13  |
| 2.2.3. Sınıflandırma Analizi.....                                           | 13  |
| 2.2.4. Karar Ağacı.....                                                     | 14  |
| 2.2.5. Random Forest .....                                                  | 17  |
| 2.2.6. Destek Vektör Makineleri.....                                        | 17  |
| 2.2.7. Lojistik Regresyon .....                                             | 18  |
| 2.3. Sınıflandırıcı Değerlendirme Ölçüleri .....                            | 18  |
| 2.3.1. Doğruluk .....                                                       | 18  |
| 2.3.2. Karışıklık matrisi.....                                              | 18  |
| 2.3.3. Alıcı İşletim Eğrisi (receiver operating characteristics: ROC) ..... | 21  |
| 2.3.4. Diyabet Tahmini ile İlgili Literatür Araştırması .....               | 21  |
| 2.3.5. Makine Öğrenmesi ile Yapılan Literatür Çalışmaları.....              | 25  |
| 2.3.6. Diyabette Kan Hücreleriyle Yapılan Çalışmalar .....                  | 26  |
| <b>3. GEREÇ ve YÖNTEM</b> .....                                             | 29  |
| 3.1. Veri toplama .....                                                     | 29  |
| 3.2. Makine öğrenimi araç seti .....                                        | 31  |
| 3.3. Veri kümesi açıklaması.....                                            | 31  |
| 3.4. Veri ön işleme.....                                                    | 33  |

|                                   |    |
|-----------------------------------|----|
| <b>4. BULGULAR</b> .....          | 34 |
| 4.1. Lojistik Regresyon .....     | 35 |
| 4.2. Karar Ağacı.....             | 38 |
| 4.3. Rastgele Orman .....         | 40 |
| <b>5. TARTIŞMA</b> .....          | 43 |
| 5.1 Kısıtlılıklar .....           | 44 |
| <b>6. SONUÇ VE ÖNERİLER</b> ..... | 46 |
| <b>KAYNAKLAR</b> .....            | 48 |



## TABLÖLAR DİZİNİ

|                                                                                 |     |
|---------------------------------------------------------------------------------|-----|
| <b>TABLO 2. 1</b> DİYABETLİ KİŞİ SAYISI EN YÜKSEK İLK 5 ÜLKE .....              | 10  |
| <b>TABLO 2. 2</b> ÜLKELERE GÖRE 20-79 YAŞ GRUBU DİYABETLİ SAYISI .....          | 10  |
| <b>TABLO 3. 1</b> VERİ SETİNDE KULLANILAN ÖZELLİKLER VE TİPLERİ .....           | 30  |
| <b>TABLO 3. 2</b> VERİ KÜMESİNİN İSTATİSTİKSEL ÖLÇÜMLERİ .....                  | 32  |
| <b>TABLO 4. 1</b> KULLANILAN VERİ KÜMESİNİN İSTATİSTİKSEL ÖLÇÜMLERİ.....        | 35  |
| <b>TABLO 4. 2</b> TEK DEĞİŞKENLİ ANALİZ ÇALIŞMASI SONUCU ÖZET TABLOSU .....     | 35  |
| <b>TABLO 4. 3</b> ÇOK DEĞİŞKENLİ ANALİZ ÇALIŞMASI SONUCU ÖZET TABLOSU .....     | 36  |
| <b>TABLO 4. 4</b> TEK DEĞİŞKENLİ VE ÇOK DEĞİŞKENLİ ANALİZLER ÖZET TABLOSU ..... | 37  |
| <b>TABLO 4. 5</b> LOJİSTİK REGRESYON SINIFLANDIRMASI KARIŞIKLIK MATRİSİ .....   | 38  |
| <b>TABLO 4. 6</b> KARAR AĞACI SINIFLANDIRMASI KARIŞIKLIK MATRİSİ .....          | 38  |
| <b>TABLO 4. 7</b> RANDOM FOREST SINIFLANDIRMASI KARIŞIKLIK MATRİSİ .....        | 40  |
| <b>TABLO 5. 1</b> MAKİNE ÖĞRENMELERİ PERFORMANS ÖLÇÜMLERİ KARŞILAŞTIRMASI ..... | 449 |

## ŞEKİLLER DİZİNİ

|                                                                                    |    |
|------------------------------------------------------------------------------------|----|
| ŞEKİL 2. 1 ERKEKLER VE KADINLARDAKİ DİYABET HASTALIĞININ YAYGINLIĞI .....          | 19 |
| ŞEKİL 2. 2 KARIŞIKLIK MATRİSİ .....                                                | 19 |
| ŞEKİL 3. 1 VERİLERİN TOPLANDIĞI ARAYÜZ .....                                       | 31 |
| ŞEKİL 3. 2 ARAŞTIRMA SÜRECİ ÖNERİLEN ÇERÇEVE .....                                 | 32 |
| ŞEKİL 4. 1 MAKİNE ÖĞRENMELEİ ACCURACY ÖLÇÜMLERİ .....                              | 37 |
| ŞEKİL 4. 2 LOJİSTİK REGRESYON ROC AUC .....                                        | 38 |
| ŞEKİL 4. 3 DİYABETLİLERLE SAĞLIKLI LARİ AYIRMAK İÇİN OLUŞTURULAN KARAR AĞACI ..... | 39 |
| ŞEKİL 4. 4 KARAR AĞACI ROC AUC .....                                               | 39 |
| ŞEKİL 4. 5 RASTGELE ORMAN ROC AUC .....                                            | 40 |
| ŞEKİL 4. 6 CİNSİYETLERE GÖRE YAŞ DAĞILIM GRAFİĞİ .....                             | 41 |
| ŞEKİL 5. 1 ALGORİTMALARIN İSTATİKSEL KARŞILAŞTIRILMASI .....                       | 44 |



## SİMGELER ve KISALTMALAR

|               |                                                              |
|---------------|--------------------------------------------------------------|
| <b>AB</b>     | : Adaboost                                                   |
| <b>ACC</b>    | : Accuracy (Kesinlik)                                        |
| <b>ADA</b>    | : American Diabetes Association (Amerikan Diyabet Derneği)   |
| <b>AI</b>     | : Artificial Intelligence (Yapay Zeka)                       |
| <b>ALT</b>    | : Alanin Aminotransferaz                                     |
| <b>ANN</b>    | : Artificial Neural Network (Yapay Sinir Ağı)                |
| <b>AUC</b>    | : Area Under the ROC Curve                                   |
| <b>BN</b>     | : Bayes Network                                              |
| <b>CART</b>   | : Classification and Regression Trees                        |
| <b>CRP</b>    | : c-Reaktif Protein                                          |
| <b>DL</b>     | : Deep Learning (Derin Öğrenme)                              |
| <b>DM</b>     | : Diabetes Mellitus                                          |
| <b>DT</b>     | : Desicion Tree (Karar Ağacı)                                |
| <b>DVM</b>    | : Destek Vektör Makinesi                                     |
| <b>GDM</b>    | : Gestasyonel Diyabetes Mellitus                             |
| <b>G-GT</b>   | : Gama-Glutamil Transferaz                                   |
| <b>HGB</b>    | : Hemoglobin                                                 |
| <b>HDL</b>    | : High Density Lipoprotein                                   |
| <b>ICD-10</b> | : International Classification of Diseases (ICD) Revision-10 |
| <b>ID3</b>    | : Iterative Dichotomiser 3                                   |

|              |                                                                              |
|--------------|------------------------------------------------------------------------------|
| <b>IDDM</b>  | : Insulin Dependent Diabetes Mellitus                                        |
| <b>IDF</b>   | : International Diabetes Federation                                          |
| <b>KA</b>    | : Karar Ağacı                                                                |
| <b>k-NN</b>  | : k-Nearest Neighbor                                                         |
| <b>LADA</b>  | : Yetişkinlerde Gizli Otoimmün Diyabet)                                      |
| <b>LDL</b>   | : Low Density Lipoprotein                                                    |
| <b>LR</b>    | : Logistic Regression                                                        |
| <b>l-SVM</b> | : Linear-Support Vector Machine                                              |
| <b>MCH</b>   | :Mean Corpuscular Hemoglobin                                                 |
| <b>MCHC</b>  | : Mean Corpuscular Hemoglobin Concentration                                  |
| <b>ML</b>    | : Machine Learning (Makine Öğrenmesi)                                        |
| <b>MLP</b>   | : Multilayer Perceptron                                                      |
| <b>MODY</b>  | : Maturity-Onset Diabetes of Young (Gençlerin Olgunluk Başlangıçlı Diyabeti) |
| <b>MÖ</b>    | : Makine Öğrenmesi                                                           |
| <b>MPV</b>   | : Mean Platelet Volume                                                       |
| <b>NA</b>    | : Sodyum                                                                     |
| <b>NB</b>    | : Naive Bayes                                                                |
| <b>OGTT</b>  | : Oral Glukoz Tolerans Testi                                                 |
| <b>PCA</b>   | : Principal Component Analysis (Temel Bileşen Analizi)                       |
| <b>PCT</b>   | : Prokalsitonin                                                              |
| <b>PLT</b>   | : Trombosit                                                                  |

|             |                                                                |
|-------------|----------------------------------------------------------------|
| <b>RA</b>   | : Regresyon Ağacı                                              |
| <b>RBC</b>  | : Eritrosit                                                    |
| <b>RDW</b>  | : Eritrosit Dağılım Genişliği                                  |
| <b>RF</b>   | : Random Forest                                                |
| <b>RO</b>   | : Rastgele Ormanlar                                            |
| <b>ROC</b>  | : Receiver Operating Characteristic                            |
| <b>SGPT</b> | : Alanin aminotransferaz                                       |
| <b>SL</b>   | : Supervised Learning (Gözetimli Öğrenme)                      |
| <b>SVM</b>  | : Support Vector Machine (Destek Vektör Makinesi)              |
| <b>T1DM</b> | : Tip 1 Diyabetes Mellitus                                     |
| <b>T2DM</b> | : Tip 2 Diyabetes Mellitus                                     |
| <b>UACR</b> | : Urine Albumin/Creatinin Rate (İdrar Albümin/Kreatinin Oranı) |
| <b>u-SL</b> | : Unsupervised Learning (Gözetimsiz Öğrenme)                   |
| <b>VKİ</b>  | : Vücut Kütle İndeksi                                          |
| <b>YSA</b>  | : Yapay Sinir Ağı                                              |
| <b>WHO</b>  | : World Health Organization (Dünya Sağlık Örgütü)              |

## 1. GİRİŞ

Uluslararası Diyabet Federasyonu 2021 verilerine göre, her 5 saniyede bir kişi diyabet sebebiyle ölüyor. Dünyada 537 milyon diyabet hastasının bulunduğunu ve gerekli önlemler alınmadığı takdirde bu sayının 2030 yılı tahminlerinde 643 milyona, 2045 yılı tahminlerinde ise 783 milyon diyabet hastasına çıkacağı ön görülmüştür.

Diyabetin manevi boyutu ile beraber maddi boyutu da göz önüne alındığında Diyabet Atlası 2021 verilerine göre küresel sağlık harcamalarının yaklaşık %11.5'lik kısmı (yaklaşık 966 milyar \$) diyabet hastalığı teşhis ve tedavisi için kullanılmıştır. Son 15 yıldaki harcamalara bakıldığında ise %316 artışa karşılık geldiği ve 2030 yılına kadar hasta oranlarının daha da artacağı görülmüştür.

2021'de diyabetli (20-79 yaş arası) neredeyse iki yetiştikten birinin (%44,7; 239,7 milyon) durumlarından habersiz olduğu bulunmuştur. Komplikasyonları önlemek veya geciktirmek, erken ölümden kaçınmak ve yaşam kalitesini iyileştirmek için diyabetli kişilerin mümkün olduğunca erken teşhis edilmesinin önemli olduğu görülmektedir. Ayrıca diyabet tanısı almış kişiler sağlık sistemlerine ek bir yük oluşturmaktadır. Diyabetli olduklarının farkında olmayan diyabetli kişileri teşhis etme ve diyabetli tüm insanlara mümkün olan en erken zamanda uygun ve zamanında bakım sağlama becerisini iyileştirmeye yönelik acil ihtiyacı vurgulamaktadır (IDF, 2021).

Bu çalışmanın amacı yapay zeka teknolojilerinden yararlanılarak hastalara ait tam kan sayımı sonuçlarından tip 2 diyabet hastalığının önceden öngörülmesi konusunun ele alınmasıdır. Böylelikle erken tanı ile beraber sağlık sistemlerindeki ek iş yükü hafifleyecek, hastaların yaşam kalitesi iyileşecek ve yüksek tedavi giderleri açısından ekonomik olarak sıkıntıya düşmesi engellenecektir. Kan bazlı değişkenlerin, gelecekteki tip 2 diyabet riskinin ölçülmesinde ve hastalığa olası etiyolojik yolların anlaşılmasında literatüre katkıda bulunacaktır.

## 2. GENEL BİLGİLER

### 2.1. Diabetes Mellitus

#### 2.1.1. Diyabetin Tarihçesi

Antik çağda milattan önce 1500'e kadar uzanan yıllarda siyah karıncalar diyabeti binlerce yıl önceden tespit etmişlerdir. Karıncalar ve sinekler kurbanların idrarının bulunduğu yere doğru toplanarak, ilk kayıtlarda günümüzdeki şeker hastalığına benzerlik gösteren aşırı idrar çıkışı ve yoğun susuzluğa neden olan gizemli ve ölümcül bir hastalıktan bahsedilmiştir (J. Maccracken, 1997).

Milattan sonra birinci yüzyılda ise Kapadokya'da Arateus tarafından ilk tanım Yunancada "sifon" veya "geçmek", hastaların tüketebileceklerinden daha fazla sıvı tüketmesi anlamına gelen diyabet kelimesi ile tanımlanmıştır (Lakhtakia, 2013).

1869'da Alman tıp öğrencisi Paul Langerhans pankreasta adacık hücreleri bulmuş, ancak işlevlerini açıklayamamıştır. Yirmi yıl sonra, Joseph von Mehring ve Oskar Minkowski, köpeklerin pankreasını çıkardıklarında diyabete yol açtığını öğrenmişlerdir (Torino, 1989).

Banting ve Best, 1921'de köpeklerin pankreas kanallarını bağlamışlardır; birkaç hafta sonra, atrofik pankreası çıkarmışlardır ve pankreası birkaç gün önce alınmış olan Marjorie adlı hasta bir köpeğin damarına enjekte ettikleri bir özü oluşturmuşlardır. Köpeğin kan şekeri seviyelerinde bir düşüş olmuştur. Ancak kan şekerindeki düşüş geçici olduğu için bu başarı kısa sürmüştür. Ekip, bir biyokimyacı olan Doktor James Collip'in yardımıyla sonunda insan kullanımı için insülini saflaştırmayı başararak tarihte bir yer edinmişlerdir (Siang Yong Tan, 2017).

Collip, 23 Ocak 1922 tarihinde insulini daha da saflaştırarak bir diyabet hastası olan Leonard Thompson üstünde ilk defa denemiş, sonucunda ise tam bir başarı elde etmiştir. İnsülinin başarısına dair haberler yayıldıkça, talep fırlamıştır. Banting ve Best, insülinin çok daha büyük miktarlarda üretilebilmesi için tekniklerini geliştirmeye başlamışlardır. Mayıs 1922'de Eli Lilly, seri insülin üreten ilk üretici olmuştur (Kenéz, 1978).

1955 de, Frederick Sanger, 14 yıllık çalışmasının ardından insülinin kimyasal yapısını tanımlamış ve insulinin iki aminoasit zinciri yapısında olduğunu bularak Nobel ödülü almıştır (Majumdar, 2001). Fredrick Banting'in doğum günü olduğu için Dünya diyabet günü her yıl 14 Kasım'da gerçekleştirilmektedir (Shahbazian, 2013).

### **2.1.2. Diyabetin Tipleri**

Diyabetli kişilerde kan şekeri seviyeleri yüksek kalır. Bunun nedeni insülinin hiç üretilmemesi, yeterli düzeyde yapılmaması veya olması gerektiği kadar etkili olmaması olabilir. Diyabetin en yaygın biçimleri otoimmün hastalığı olarak kabul gören tip 1 diyabet (%5) ve fazla kilo ile ilişkili kabul gören tip 2 diyabet (%95). Gestasyonel diyabet, gebelikte yaşanan bir diyabet türüdür. Diğer diyabet türleri çok nadirdir ve tek gen mutasyonundan kaynaklanmaktadır (Ozougwu, 2013).

#### **Tip 1 Diyabet**

İnsülinin üretildiği pankreatik  $\beta$  hücrelerinde oluşan seçici yıkımı ile ilişkili kronik otoimmün bir hastalıktır (Ozougwu, 2013). İkiz donörlerden kronik diyabetik ikiz alıcılara pankreas transplantasyonu yapıldığında,  $\beta$ -hücrelerinin pankreas lezyonlarının yüksek heterojenliği nedeniyle immün baskılanma olmadan komplike hale gelmekte ve daha sonra infiltre edici Tlenfositler kullanılarak masif insülin gelişmektedir. Sık idrara çıkma, susuzluk, kilo kaybı, aşırı yorgunluk, aseton nefesi, mide bulantısı ve kusma, bulanık görme ve genital bölgede kaşıntı genel semptomlar arasındadır (Deepthi B, 2017).

#### **Tip 2 Diyabet**

Bu diyabet türü yetişkinlikten itibaren başlayan bir diyabet türüdür. Bu tipe sahip diyabetli kişiler sıklıkla insülinin etkisine dirençlidir. Küresel olarak, dünya nüfusunun %5-7'sini etkilemektedir. Hastalık genellikle diyet tedavisi, egzersiz ve hipoglisemik ajanlarla kontrol edilmektedir. Bu diabetes mellitusun en yaygın şeklidir ve ailenin diyabet geçmişi, yaş, kilo ve spor alışkanlıkları ile yüksek oranda bir korelasyona sahiptir (Bastaki, 2005).

#### **Gestasyonel diyabet**

Hamile kadınlar sıklıkla diyabet geliştirmektedir. Hamilelik sırasında büyük miktarlarda hormon üretilir, bu hormonlar annenin vücudundaki insülin etkisini azaltarak insülin direncine neden olabilmektedir. Hamilelik esnasında gelişen ve keşfedilen tanısı konmamış asemptomatik tip 2 diyabet, gestasyonel diyabet olarak sınıflandırılmaktadır. Gestasyonel diyabetin klinik önemi, önemli maternal ve fetal morbidite ile ilişkili olması gerçeğinde yatmaktadır (Wondifraw, 2015).

### **2.1.3. Diabetes mellitus için tanı testleri**

Prediyalet ve diyabet teşhisi için dört kan testi mevcuttur. Bunlar gündelik plazma (kan) glikoz/rastgele plazma testi, oral glukoz tolerans testi (OGTT), hemoglobin A1C (HbA1c), açlık plazma glukozudur.

### **Gündelik plazma (kan) şekeri/ rastgele plazma testi**

En basit testtir ve teste girmeden önce açlığı gerektirmez. Bu test ile diyabet teşhisi için kriterler, diyabet semptomlarının varlığı ve kan şekeri seviyesinin 200mg/dl veya daha yüksek olmasıdır (Harikumar K., 2016).

### **Açlık plazma glukozu**

7,0 mmol/L'lik bir açlık plazma glukoz seviyesi, 75 g'lık bir oral glukoz tolerans testinde 2 saatlik plazma glukoz değeri  $\geq 11,1$  mmol/L ile en yakından ilişkilidir ve her biri retinopati gelişimini öngörmektedir. Bunu almadan önce 8 saat açlık olmalıdır. Farklı günlerde yapılan iki veya daha fazla testte kan şekerinin 126mg/dl'den fazla olması diyabet tanısını doğrulamaktadır (Committee & Ronald Goldenberg, 2013).

### **Oral glikoz tolerans testi**

2 saatlik kan şekeri seviyesi 200mg/dl veya daha yüksek, 2 saatlik kan şekeri seviyesi 140-199mg/dl ise prediyabet tanısı konmaktadır(Harikumar K., 2016).

### **Hemoglobin A1C (HbA1c)**

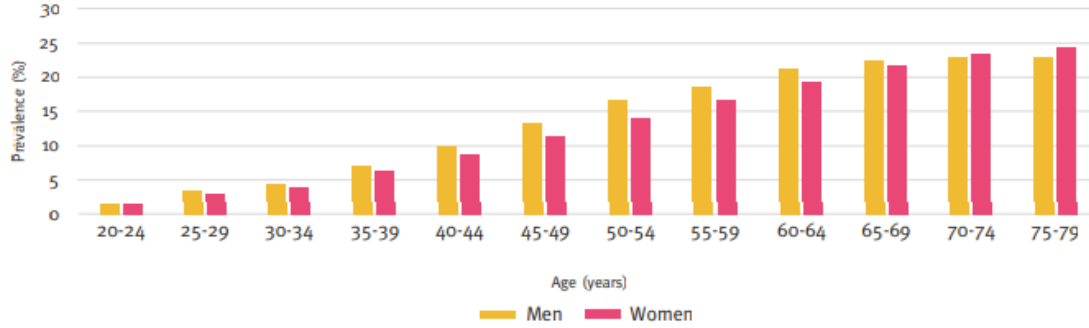
A1C günün herhangi bir saatinde ölçülebilmektedir ve 75g OGTT'de açlık plazma glukozundan veya 2 saatlik plazma glikozundan daha uygundur. A1C, önceki 2-3 aydaki ortalama plazma glikozunu yansıttığından, glikoz değerlerinin günden güne değişkenliği sorununu da ortadan kaldırmaktadır. HbA1c  $\geq 6,5$ 'in özgülüğü diyabet tanısını doğrulamak için yeterince yüksektir ve HbA1c  $\leq 5,7$ 'nin duyarlılığı diyabet tanısının dışlanmasını haklı çıkarmak için yeterince yüksektir (Committee & Ronald Goldenberg, 2013).

### **Dünyada ve Türkiye’de Diyabet**

Küresel olarak, her 10 yetişkinden biri şu anda diyabetle yaşamaktadır. Ayrıca, yetişkin nüfusun beşte birinin hatta daha fazlasının diyabetli olduğu ülkelerin sayısı giderek artmaktadır. 2000 yılından günümüze, 20 ile 79 yaş aralığındaki yetişkin bireylerde diyabet tahmini prevalans oranı üç kattan daha çok artıp 151 milyondan (dünya nüfusunun %4,6'sı) bugün 537 milyona (%10,5) çıkmıştır (IDF, 2021).

2017'de bu duruma 1 milyondan fazla ölüm ile ölüm nedenleri arasında dokuzuncu sırada yer almıştır. Bu, tip 2 diyabetin on sekizinci önde gelen ölüm nedeni olarak sıralandığı 1990 ile karşılaştırıldığında önemli bir artıştır olduğu görülmektedir (Khan, 2019). Bu sebeple diyabet, dünya genelindeki ölümlerin önemli bir kısmını oluşturmaktadır. İstatistiksel verilere bakıldığında diyabet hastalığının yaygınlığı dikkat çekici bir oranda artış göstermektedir.

Bu durumun farkına varan tüm dünya ülkeleri, diyabet hastalığındaki ilerlemenin durdurulması konusunda hemfikir olmuş ve 2025 yılı itibarıyla hastalığın önemli bir ölçüde önüne geçilmesini kararlaştırmışlardır (World Health Organization, 2022).



**Şekil 2.1** Erkekler ve kadınlardaki diyabet hastalığının yaygınlığı

IDF, diyabete bağlı doğrudan küresel sağlık harcamalarının bir trilyon dolara yakın olduğunu ve 2030 yılına kadar bu rakamı aşacağını açıklamıştır. IDF, küresel diyabet harcamalarının 2030 yılına kadar 825 milyar ABD dolarına ve 2045 yılına kadar 845 milyar ABD dolarına ulaşacağını açıklamıştır. Uluslararası Diyabet Federasyonu'na göre Türkiye, 2021 yılında DM'li bir bireyin tedavisi için ortalama 1404 dolar harcamaktadır (IDF, 2021). Türkiye Diyabet Vakfına göre, Türkiye'de 10 milyon diyabet hastası olduğu tahmin edilmektedir (Türkiye Diyabet Vakfı, 2022).

**Tablo 2. 1** Diyabetli Kişi Sayısı En Yüksek İlk 5 Ülke

| Diyabetli kişi sayısı en yüksek ilk 5 ülke (20-79 yaş) (milyon) |        |       |
|-----------------------------------------------------------------|--------|-------|
|                                                                 | 2011   | 2021  |
| Türkiye                                                         | 3.5 m  | 9 m   |
| Rusya                                                           | 12.6 m | 7.4 m |
| Almanya                                                         | 5 m    | 6.2 m |
| İspanya                                                         | 2.8 m  | 5.1m  |
| İtalya                                                          | 3.6 m  | 4.5 m |

Uluslararası Diyabet Vakfı verilerine göre Avrupa kıtasındaki 59 ülke arasında son yıllık verilere bakıldığında 2011 yılı 20-79 yaş grubu diyabet prevalansının %8.1 den 2021 yılı 20-79 yaş grubu diyabet prevalansının %14.5'e çıkarak Türkiye'nin Avrupa ülkeleri arasında birinci sırayı aldığı görülmektedir. Dünya genelinde ülkemizin diyabetli kişi sayısı bakımından 2045'de 20-79 yaş grubu diyabetli bireylerde 13,4 milyon ile 10. sırada yer alacağı tahmin edilmektedir (IDF, 2021).

**Tablo 2. 2 Ülkelere Göre 20-79 Yaş Grubu Diyabetli Sayısı**

| 2021 |           |                  | 2045 |           |                  |
|------|-----------|------------------|------|-----------|------------------|
| Sıra | Ülke      | Diyabetli Sayısı | Sıra | Ülke      | Diyabetli Sayısı |
| 1    | Çin       | 140,9            | 1    | Çin       | 174,4            |
| 2    | Hindistan | 74,2             | 2    | Hindistan | 124,9            |
| 3    | Pakistan  | 33,0             | 3    | Pakistan  | 62,2             |
| 4    | ABD       | 32,2             | 4    | ABD       | 36,3             |
| 5    | Endonezya | 19,5             | 5    | Endonezya | 28,6             |
| 6    | Brezilya  | 15,7             | 6    | Brezilya  | 23,2             |
| 7    | Meksika   | 14,1             | 7    | Bangladeş | 22,3             |
| 8    | Bangladeş | 13,1             | 8    | Meksika   | 21,2             |
| 9    | Japonya   | 11,0             | 9    | Mısır     | 20,0             |
| 10   | Mısır     | 10,9             | 10   | Türkiye   | 13,4             |

Sosyoekonomik durumun diyabete etkisini belirlemek için yapılan bir çalışmada, Avrupa’da ve İskandinavya’daki gelişmiş ülkelerin ekonomik ve sağlıksal durumlarına bakıldığında başarı oranının epey yüksek olduğu görülmüştür. Bu ülkelerdeki kişi başına düşen gelir ile diyabet prevalansı arasında ters orantılı bir ilişki olduğu saptanmıştır (Oguzhan, 2020). Son senelerde yapılan çalışmalarda, yaşam tarzının da daha çok ilgi görmeye başladığı ve sağlıkta değerli bir etken olduğu görülmüştür (FARHUD, 2015). Sağlıklı bir yaşam tarzı sağlamak için Dünya Sağlık Örgütü, bol meyve ve sebze yemeyi, yağ, şeker ve tuz alımını azaltmayı ve egzersiz yapmayı önermiştir. Boy ve kilo bazında, insanların aşırı kilolu olup olmadıklarını görmek için vücut kitle indekslerini (VKİ) kontrol ettirmesini tavsiye etmişlerdir (WHO, 2010).

Bugün Türkiye’de beslenme şeklimizden kaynaklı sorunlar gelişmekte olan ve gelişmiş ülkeler ile karşılaştırıldığında aynı görünüme sahip olduğu görülmüştür. Halkımızın temel besinlerine bakıldığında ekmek ve tahıl ürünleri olduğu görülmektedir. Yıllar içerisindeki besin tüketimi incelendiğinde et ve süt ürünleri, meyve ve taze sebze tüketiminin azaldığı; kuru baklagiller, yumurta ve şeker tüketiminin arttığı gözlenmiştir (Türkiye’de Diyabette Çözüm Yolları, 2020).

## 2.2. Yapay Zeka ve Makine Öğrenmesi

Yapay zeka, ilk olarak robotlar ile karşımıza çıkmıştır. Çek dilinde “robota” olarak adlandırılan bu kelime, literatürde 1921 senesinde Karel Capek’in yazmış olduğu “Rossum’un Evrensel Robotları” adlı kitabında biyosentetik makinelerin zorla çalıştırıldığı bir fabrika anlamına gelmiştir (wikipedia, 2023).

“Yapay Zeka”nın kurucu babalarından biri olan bilgisayar bilimci John McCarthy, çöp toplama yöntemini icat etmiş ve ilk yapay zekanın geliştirilmesinde çok büyük bir katkısı olmuştur. McCarthy aynı zamanda 1961 yılında yardımcı bilişim fikrini bulmuştur. Bilgisayar veya bilgi yardımcı programı fikri, 1970’lere doğru popüler olmuştur. Fakat 1990 ların ortalarına doğru ve 2000 li yıllarda bu fikir yepyeni bir şekle evrilmiştir (wikipedia, 2023).

Bugün yapay zeka kavramı, komplike sorunları aşmak adına öneriler üreten ve bunları uygulayıp test eden bir mühendislik alanı olarak geçerliliği devam etmiştir (Needham, 1959). Yapay zekanın tıpta uygulanmasının iki ana dalı vardır. Bunlar sanal ve fiziksel olarak adlandırılmıştır. Sanal bileşen, deneyim yoluyla öğrenmeyi geliştiren matematiksel algoritmalarla temsil edilen Makine Öğrenimi ile temsil edilmiştir. Makine Öğrenimi algoritmaları temel olarak dört kategoriye ayrılmıştır. Bunlar denetimli öğrenme, denetimsiz öğrenme, yarı denetimli öğrenme ve takviyeli öğrenme olarak sıralanmıştır.

Denetimli öğrenme, veri bilimi ve makine öğrenmesi alanlarında kullanılan önemli bir yöntemdir. Bu yöntem, verilerin sınıflandırılması veya modellemesi gibi belirli bir amaç doğrultusunda verilerin öğrenilmesini ve tahmin yapılmasını sağlamaktadır. Denetimli öğrenme yöntemleri, sınıflandırma ve gerileme problemleri gibi iki temel problem türünü çözebilmektedir. Sınıflandırma problemi, veri kümesindeki her bir örneği belirli bir sınıfa atama işlemidir. Gerileme problemi ise veri kümesindeki veriler arasındaki ilişkiyi modelleme işlemidir. Modelin performansı verilerin doğruluğu ve yöntem seçimine bağlıdır (M. Razzak, 2018).

Denetimsiz öğrenme, verilerin sınıflandırılması, etiketlenmesi veya yapılandırılması olmadan verilerin öğrenilmesini ve keşfedilmesini amaçlamaktadır. En yaygın denetimsiz öğrenme görevleri arasında, kümeleme, boyutsallık azaltma ve üretici modeller vb. sayılabilir. Verilerin sınıflandırılması ve etiketlenmemiş olması gibi avantajlarına rağmen, verilerin anlamının yanlış yorumlanması ve veriler arasındaki bağlantıların belirlenmesi gibi dezavantajları da bulunmaktadır (Mittal, 2018).

Yarı denetimli öğrenme, hem etiketli hem de etiketsiz veriler üzerinde çalıştığı için yukarıda belirtilen denetimli ve denetimsiz yöntemlerin bir melezlenmesi olarak tanımlanabilir. Dolayısıyla “denetimsiz” öğrenme ile “denetimli” öğrenme arasında kalmaktadır. Yarı denetimli bir öğrenme modelinin nihai amacı, modelden yalnızca etiketlenmiş veriler kullanılarak üretilenden daha iyi bir tahmin sonucu sağlamaktır. Araştırmalara göre, yarı denetimli öğrenme, verilerin içeriğinin ve özelliklerinin zor olduğu ve etiketli verilerin sınırlı olduğu problemlerde etkilidir. Yarı denetimli öğrenmenin kullanıldığı bazı uygulama alanları arasında makine çevirisi, dolandırıcılık tespiti, etiketleme verileri ve metin sınıflandırması bulunmaktadır. (Ratray, 2007) (Mittal M. R., 2018)

Takviye öğrenimi, yazılım araçlarının ve makinelerin verimliliğini artırmak için belirli bir bağlamda veya ortamda en uygun davranışı otomatik olarak değerlendirmesini sağlayan bir tür makine öğrenimi algoritmasıdır, yani çevre odaklı bir yaklaşımdır. Bu tür öğrenme, ödül veya cezaya dayalıdır ve nihai amacı, ödülü artırmak veya riski en aza indirmek için harekete geçmek için çevre aktivistlerinden elde edilen bilgileri kullanmaktır (Sarker, 2021).

### **2.2.1. Makine Öğrenmesi**

Makine Öğrenimi, Bilgi Teknolojisinin temel dayanaklarından biri haline gelmiş ve bununla birlikte, genellikle gizli olsa da, hayatımızın oldukça merkezi bir parçası haline gelmiştir. En önemlisi veri madenciliği olan Makine Öğrenimi (ML) için çeşitli uygulamalar vardır. İnsanlar genellikle analizler sırasında veya muhtemelen birden fazla özellik arasında ilişki kurmaya çalışırken hata yapmaya eğilimlidir (Akinsola, 2017).

ML, kalıpları ortaya çıkarmak ve davranışları tahmin etmek için veriye dayalı teknikleri kullanan bilgisayar biliminde yapay zekanın bir alt alanıdır. Son yıllarda, hastalık teşhisi, tedavisi ve prognozunda doğru ve etkili olduğu kanıtlanmış olan makine öğrenimi teknikleri tıp ve sağlık alanında yaygın olarak uygulanmaktadır (Topol, 2019). Makine öğrenimi algoritmalarına dayalı tahmine dayalı modeller, diyabetli hastaların belirlenmesinde faydalı olabileceği ve diyabet risk faktörlerinde gözden kaçabilecek gizli kalıpları keşfetmeye yardımcı olabilir olduğu görülmüştür (Lai H, 2019).

### **2.2.2. Makine Öğrenimi Görevleri ve Algoritmaları**

Makine öğrenimi tabanlı bir tahmine dayalı modelin genel yapısı, modelin 1. aşamada geçmiş verilerden eğitildiği ve 2. aşamada yeni test verileri için sonucu üretilmiştir.

### **2.2.3. Sınıflandırma Analizi**

Sınıflandırma, belirli bir örnek için bir sınıf etiketinin tahmin edildiği bir tahmine dayalı modelleme sorununa atıfta bulunarak, makine öğreniminde denetimli bir öğrenme yöntemi olarak kabul edilmiştir. Matematiksel olarak, bir fonksiyonu ( $f$ ) girdi değişkenlerinden ( $X$ ) çıktı değişkenlerine ( $Y$ ) hedef, etiket veya kategoriler olarak eşlemektedir. Verilen veri noktalarının sınıfını tahmin etmek için yapılandırılmış veya yapılandırılmamış veriler üzerinde gerçekleştirilebilmektedir.

## **İkili Sınıflandırma**

“Doğru ve yanlış” veya “evet ve hayır” gibi iki sınıf etiketine sahip sınıflandırma görevlerini ifade etmektedir. Bu tür ikili sınıflandırma görevlerinde, bir sınıf normal durum olabilirken, anormal durum başka bir sınıf olabilmektedir. Örneğin, "kanser tespit edilmedi", tıbbi test içeren bir görevin normal durumudur ve "kanser tespit edildi", anormal durum olarak kabul edilmiştir (Islam R. A., 2019).

## **Çok Sınıflı Sınıflandırma**

Geleneksel olarak bu, ikiden fazla sınıf etiketine sahip olan sınıflandırma görevlerini ifade etmektedir. Çok sınıflı sınıflandırma, ikili sınıflandırma görevlerinden farklı olarak normal ve anormal sonuçlar ilkesine sahip değildir. Bunun yerine, belirli bir sınıf aralığı içinde örnekler birine ait olarak sınıflandırılmıştır (Jiawei Han, 2012).

## **Çok Etiketli Sınıflandırma**

Makine öğreniminde, bir örneğin birkaç sınıf veya etiketle ilişkilendirildiği durumlarda çok etiketli sınıflandırmanın önemli olduğu kabul edilmiştir. Bu nedenle, probleme dahil olan sınıfların hiyerarşik olarak yapılandırıldığı ve her örneğin aynı anda her hiyerarşik seviyede birden fazla sınıfa, örneğin çok seviyeli metin sınıflandırmasına ait olabileceği çok sınıflı sınıflandırmanın bir genellemesidir. Çok etiketli sınıflandırma, sınıf etiketlerinin birbirini dışladığı geleneksel sınıflandırma görevlerinden farklı olarak, karşılıklı olarak özel olmayan çeşitli sınıfları veya etiketleri tahmin etmeyi destekleyen gelişmiş makine öğrenimi algoritmalarını içermektedir (Pedregosa, 2012). Bu bölüm, hastalık teşhisinde en sık kullanılan makine öğrenimi algoritmalarının kapsamlı bir incelemesini sağlamaktadır.

### **2.2.4. Karar Ağacı**

Karar ağacı çıkarımı, örüntü sınıflandırması için çok popüler ve pratik bir yaklaşımdır. Karar ağacı tümevarımı, sınıf etiketli eğitim gruplarından karar ağaçlarının öğrenilmesidir.

Karar ağacının başlıca iki türü vardır:

Tahmini edilmiş sonucun verilerinin ait olduğu sınıf olduğunda analizin bulunduğu sınıflandırma ağacı (Classification Tree).

Tahmini edilmiş sonucun gerçek bir sayı olarak kabulü edilebilecek analizin olduğu regresyon ağacı (Regression tree).

Hem sınıflandırma hem de regresyon ağaçları, nerede bölüneceğini belirlemek için kullanılan prosedür gibi farklılıkların yanı sıra benzerliklere de sahiptir. ID3, C4.5, CART, CHAID, MARS olmak üzere çeşitli karar ağaçları algoritmaları vardır (Verma, 2017).

### Entropi

Rastgele bir değişkenle ilişkili belirsizliğin bir ölçüsüdür. Entropi, belirsizlik veya rastgelelikteki artışla artmakta ve belirsizlik veya rastgelelikteki azalmayla azalmaktadır. Entropi değeri 0-1 arasında değişmektedir.

$$Entropi(D) = \sum_i^c = 1 - p_i \log_2(p_i)$$

Burada  $p_i$ , H'deki rastgele bir tanımlama grubunun C sınıfına ait olması ve  $|C_i, D|/|D|$  ile tahmin edilmesinin sıfır olmayan olasılığıdır. Entropi yukarıda belirtildiği gibi, taban 2'nin bir günlük işlevi kullanılmaktadır. 0 ve 1 bitlerinde kodlanmıştır (Kamber, 2011).

### Bilgi Kazancı

ID3, öznitelik seçim ölçüsü olarak bilgi kazancını kullanır. Claude Shannon, mesajların değerini veya “bilgi içeriğini” incelemiş ve Bilgi Teorisinde bir ölçü olarak bilgi kazanımını vermiştir. Bilgi Kazancı, gerçek bilgi kazanımı ihtiyacı ile yeni ihtiyaç arasındaki farktır.

$$Kazanç(D, A) = Entropi(D) - \sum_{j=1}^v \frac{|D_j|}{|D|} Entropi(D_j)$$

D: Belirli bir veri bölümü

A: Özellik

V: D'deki demetleri, v farklı değerlere sahip bazı A öznitelğine böldüğümüzü varsayalım.

D, V bölümüne veya alt kümelerine bölünür,  $\{D_1, D_2, \dots, D_j\}$ ,  $D_j$ , D'de A'nın aj'sine sahip olan demetleri içeriyorsa en yüksek bilgi kazancına sahip öznitelik seçilmektedir (S. S. Al-Sarawi, 2012).

### Kazanç Oranı

Bilgi kazanımı ölçüsü, birden fazla sonuca sahip testlerde tercih edilmektedir. Her bölüm saf olduğundan, bölümlene yoluyla bilgi kazancı maksimumdur. Ancak böyle bir bölümlene sınıflandırma için kullanılamamaktadır. C4.5 bu öznitelik seçim ölçüsünü kullanmaktadır. Bilgi kazancının bir uzantısı olan Kazanç Oranı olarak adlandırılmaktadır (Hasdina, 2020).

Kazanç Oranı, bazı bölümlere göre elde edilen bir sınıflandırmaya göre bilgiyi ölçen bilgi kazancından farklıdır. Kazanç Oranı, şu şekilde tanımlanan bir "bölünmüş bilgi" değerini kullanarak bir tür bilgi kazancı uygular:

$$Bölünmüş\ Bilgi_A = - \sum_{j=1}^v \frac{|D_j|}{|D|} \log_2 \left( \frac{|D_j|}{|D|} \right)$$

Kazanç Oranı daha sonra şu şekilde tanımlanmaktadır:

$$Bilgi\ Kazancı(A) = \frac{Kazanç(A)}{Bölünmüş\ Bilgi_A(D)}$$

Maksimum Kazanç Oranına sahip nitelik olan bir bölme niteliği seçilmektedir. Bölünmüş bilgi 0 eğilimindeyse kazanç oranı kararsız hale gelmektedir. Seçilen testin bilgi kazancının en az incelenen tüm testlerdeki ortalama kazanç kadar büyük olması gerektiği gibi bir koşuldan kaçınmak için bir kısıtlama eklenmektedir (Jiawei Han, 2012).

### Gini İndeks

Gini İndeksi, CART karar ağacı algoritması tarafından kullanılan bir nitelik seçim ölçüsüdür. Gini İndeksi, D safsızlığı, bir veri bölümü veya eğitim demetleri kümesini şu şekilde ölçmektedir:

$$Gini(D) = 1 - \sum_{i=1}^m p_i^2$$

Burada  $p_i$ , D'deki bir demetin  $C_i$  sınıfına ait olma olasılığıdır ve  $|C_i, D|/|D|$  ile tahmin edilmektedir. Toplam, m sınıf üzerinden hesaplanmıştır. Safsızlığı maksimum düzeye indiren (veya minimum gini indeksine sahip olan) öznitelik, bölme özniteliği olarak seçilmiştir (Hong Han, 2016).

Karar ağacı algoritması, böl ve yönet kurallarını takip etmektedir. KA modellerinde öznitelik, sınıflandırma ağaçları olarak bilinen çeşitli değerleri alabilmektedir; yapraklar farklı sınıfları belirtirken, dallar bu sınıf etiketleriyle sonuçlanan özelliklerin kombinasyonunu yansıtmaktadır. Öte yandan KA, regresyon ağaçları adı verilen sürekli değişkenleri alabilmektedir. Verilerdeki küçük bir değişiklik, nihai tahminler üzerinde önemli bir etkiye sahip olabilmektedir. Ağacın artan karmaşıklığı nedeniyle büyük veri kümeleriyle iyi performans gösterememektedir. C4.5 ve EC4.5, iki ünlü ve en yaygın olarak kullanılan KA algoritmalarıdır (Nurrohman, 2020) (Rani, 2011).

### 2.2.5. Random Forest

Rastgele orman sınıflandırıcısında, her sınıflandırıcının girdisinden bağımsız olarak örneklenmiş rastgele bir vektör tercih edilerek üretildiği ve her ağacın bir girdi vektörünü sınıflandırmak için popülaritesi en yüksek olan sınıfı için bir birim oy kullandığı ağaç sınıflandırıcılarının bir kombinasyonundan meydana gelmiştir (Pal, 2005).

Herhangi bir örnek, ormandaki tüm ağaç tahmin edicilerinden en popüler oylanan sınıf alınarak sınıflandırılmıştır. Karar ağacı tasarımı, öznitelik tercih ölçüsünün ve bir budama yönteminin tercihi gerektirmiştir. Karar ağacı tümevarımı için kullanılan özniteliklerin seçimine yönelik birçok yaklaşım vardır ve çoğu yaklaşım özniteliğe doğrudan bir kalite ölçüsü atar. Karar ağacı tümevarımında en çok kullanılan öznitelik seçim ölçüleri Bilgi Kazanç Oranı kriteri ve Gini İndeksidir.

Rastgele orman sınıflandırıcısı, bir özniteliğin sınıflara göre ne kadar saf olduğunu ölçen bir öznitelik tercih ölçüsü olarak Gini Dizini'ni seçer. Her defasında ağaç, özelliklerin birleşimi kullanılarak yeni eğitim verileri üzerinde maksimum derinliğe kadar büyütülmüştür. Tamamen büyümüş ağaçlar budanmamıştır. Bu, Quinlan tarafından önerilen gibi diğer karar ağacı yöntemlerine göre rastgele orman sınıflandırıcısının en büyük avantajlarından biridir (Pal, Random forest classifier for remote sensing classification, 2005). Çalışmalar, öznitelik seçim ölçütlerinden çok, budama yöntem seçimlerinin ağaç tabanlı algoritmaların performanslarını değiştirdiğini göstermiştir (MINGERS, 1989).

### 2.2.6. Destek Vektör Makineleri

Destek vektör makineleri, öğrenme teorisinin önemli bir bölümünü oluşturmaktadır. Bilim ve mühendislikteki birçok uygulama için, özellikle sınıflandırma problemleri için çok verimlidirler. Fisher, Rosenblatt ve Vapnik'in verilerini ayırmak için sınıflandırma algoritmaları tarafından motive edilen destek vektör makineleri, polinom çekirdekleriyle Boser, Guyon ve Vapnik tarafından ve genel çekirdeklerle Cortes ve Vapnik tarafından tanıtılmıştır. O zamandan beri Destek Vektör Makineler, hakkında çeşitli pratik problemlere uygulamalar, orijinal modelin birçok varyasyonu ve bazı teorik araştırma çalışmaları yapılmıştır (Qiang Wu, 2006).

Destek Vektör Makinesinin temel motivasyonu, eğitim setindeki birkaç sınıfı, aralarındaki marjı en üst düzeye çıkaran bir yüzeyle ayırmaktır. Başka bir deyişle, DVM bir modelin genelleme yeteneğini en üst düzeye çıkarmaya izin vermektedir.

Bu, ampirik risk en aza indirgeme yöntemleri tarafından sıklıkla kullanılan felsefe olan eğitim verileri kümesindeki ortalama karesel hatayı en aza indirmek yerine, bir modelin genelleme hatasına bağlı bir sınırın en aza indirilmesine izin veren Yapısal Risk En Aza İndirgeme ilkesinin amacıdır (Jair Cervantes, 2020).

### 2.2.7. Lojistik Regresyon

Bir sigmoid işlevi kullanarak kategorik bağımlı değişkenin sonucunu olasılıksal olarak tahmin etmek için kullanılan ikili biçim açısından ikili çıktı üreten tahmine dayalı analitik için kullanılan denetimli bir ML modelidir.

$X_1, X_2, X_3, \dots, X_n$  ile temsil edilen 'N' girdi bağımsız değişkenleri olduğunu varsayıldığında Y'nin Evet (1) olasılığı ve 1-Y'nin Hayır (0) olasılığı olduğunu varsayalım (Devi, 2020).

Daha sonra nihai lojistik regresyon denklemi şu şekilde verilmektedir:

$$\text{Log} \left[ \frac{Y}{1-Y} \right] ==> Y = C + B_1X_1 + B_2X_2 + B_3X_3 + \dots + B_nX_n$$

çıktısı, ağırlık nasıl bağlandığına, ağırlık değerine ve teşvik işlevine bağlı olarak değişecektir.

### 2.3. Sınıflandırıcı Değerlendirme Ölçüleri

Sınıflandırıcı performansını değerlendirmek için çeşitli ölçüler vardır. Tek bir ölçü, sınıflandırıcı performansıyla ilgili tüm hikayeyi bize verememektedir. Sınıflandırıcının performansını değerlendirmek için bazı özel ölçüler kullanılmaktadır.

#### 2.3.1. Doğruluk

Bir algoritmanın doğruluğu, algoritmanın görünmeyen örnekleri ne kadar doğru sınıflandırdığının ölçüsüdür ve aşağıdaki formülle hesaplanabilmektedir:

$$\frac{\text{Doğru sınıflandırılmış örneklerin sayısı}}{\text{Toplam örnek sayısı}} \times 100$$

#### 2.3.2. Karışıklık matrisi

Doğruluk, performansı değerlendirmenin tek yolu değildir ve bazen sınıflandırıcının performansının daha ayrıntılı bir resmini istenebilmektedir. Bunlardan biri de Karışıklık Matrisi adlı bir tablodur ve şekil 2.13'de gösterilmektedir.

|             |         | Tahmin  |         |
|-------------|---------|---------|---------|
|             |         | Pozitif | Negatif |
| Gerçekleşen | Pozitif | TP      | FN      |
|             | Negatif | FP      | TN      |

Şekil 2. 2 Karışıklık Matrisi

FP, FN, TP ve TN kavramları şu şekilde tanımlanabilir:

- Yanlış pozitifler (FP): Veri nesnesinin gerçek sınıfı False ve çıktı değeri True değeri olduğunda tanımlanmaktadır.
- Yanlış negatifler (FN): Veri nesnesinin gerçek sınıfı True ve çıktı değeri False değeri olduğunda tanımlanmaktadır.
- Gerçek pozitifler (TP): Veri nesnesinin gerçek sınıfı True olduğunda ve tahmin de True değeri olduğunda, toplam doğru çıktı sayısı ile tanımlanmaktadır.
- Gerçek negatifler (TN): Veri nesnesinin gerçek sınıfı False olduğunda ve tahmin edilen de False değeri olduğunda, doğru çıktıların toplam sayısı ile tanımlanmaktadır (Ali, 2019).

Bir karışıklık matrisi, modelin ne kadar iyi çalıştığı hakkında faydalı bilgiler vermektedir. Bununla birlikte, öğeleri, daha fazla bilgi elde etmek için birçok performans metriğini hesaplamak için kullanılabilir. En popülerler arasında doğruluk, özgüllük, duyarlılık ve kesinlik bulunmaktadır.

**Doğruluk**, en sezgisel performans ölçüsüdür ve doğru sınıflandırılan nesnelerin sayısının değerlendirilen toplam nesne sayısına oranı olarak tanımlanmaktadır.

$$Acc = \frac{|TN| + |TP|}{|FN| + |FP| + |TN| + |TP|}$$

**Acc'nin tamamlayıcısı**, bir sınıflandırıcıyı yanlış tahmin yüzdesine göre değerlendiren hata oranıdır. Acc ve Err genel ölçülerdir ve doğrudan çok sınıflı sınıflandırma problemlerine uyarlabilmektedir.

$$Err = \frac{|FN| + |FP|}{|FN| + |FP| + |TN| + |TP|} = 1 - Acc$$

**Geri çağırma (R) ve özgüllük (Spe)** ölçümleri, ikili problemdeki her sınıf için bir sınıflandırıcının etkinliğini değerlendirmektedir. Duyarlılık veya gerçek pozitif oran olarak da bilinen hatırlama, pozitif sınıfa ait olan ve pozitif olarak doğru tahmin edilen örneklerin oranıdır.

$$R = \frac{|TP|}{|FP| + |TN|}$$

$$Spe = \frac{|TN|}{|FP| + |TN|}$$

**Kesinlik (P)**, olumlu bir tahminin doğru olma olasılığını tahmin eden bir ölçüdür. F-ölçüsünden kaynaklanan geri çağırma ile birleştirilmektedir. Sabit bir  $\beta$ , kesinlik ve geri çağırma arasındaki değiş tokuşu kontrol etmektedir. Genelde 1 olarak ayarlanmaktadır.

$$P = \frac{|TP|}{|TP| + |FP|}$$

$$F - measure = \frac{(\beta^2 + 1) \times P \times R}{\beta^2 \times P \times R}$$

(Eduardo P. Costa, 2005)

**Duyarlılık**, doğru sınıflandırılan olumlu örneklerin oranıdır.

$$Duyarlılık = \frac{Doğru Pozitif}{Doğru Pozitif + Yanlış Negatif}$$

**Özgüllük**, doğru sınıflandırılan olumsuz örneklerin oranıdır.

$$Özgüllük = \frac{Doğru Negatif}{Doğru Negatif + Yanlış Pozitif}$$

Doğruluk, özgüllük ve duyarlılık için baktığımızda fazla olması iyi olması anlamına gelmektedir, ancak hata oranına baktığımız zaman düşük olması iyi anlamına gelmektedir. İyi bir sınıflandırıcı, daha yüksek derecede hassas ve spesifik olmalıdır (Parikh R, 2008 ).

### 2.3.3. Alıcı İşletim Eğrisi (receiver operating characteristics: ROC)

ROC eğrileri, duyarlılık ve özgüllükten oluşturulmuş olmasına rağmen, karar eşiğine bağlı değildir. Bir ROC eğrisinde, olası her karar eşiği dikkate alınmaktadır. Bir ROC eğrisi, bir testin yanlış pozitif oranının (FPR) veya 1 - özgüllüğünün (yatay ekseninde çizilir) ve duyarlılığının (dikey ekseninde çizilir) grafiğidir.

Eğri üzerindeki her nokta, farklı bir karar eşiğinde duyarlılığı ve FPR'yi temsil eder. Çizilen (FPR, duyarlılık) koordinatlar, ampirik bir ROC eğrisi oluşturmak için çizgi bölümleriyle bağlantılıdır. ROC analizi, bir tanı testi, bir belirteç veya bir sınıflandırıcıyı temsil eden sürekli bir değişkenin ayırt edici performansını değerlendirmek için yaygın olarak kullanılır (Obuchowski, 2000).

### 2.3.4. Diyabet Tahmini ile İlgili Literatür Araştırması

Son yıllarda hastalık olasılığını tahmin etmek için veri madenciliği tekniğinin kullanımı giderek artan bir sıklıkta kullanılmaktadır. Araştırmacılar tarafından birçok algoritma ve araç seti oluşturulmuş ve incelenmiştir. Bunlar, bu araştırma alanının potansiyelini vurgulamıştır. Hastalık ve komplikasyonları önlemek ya da geciktirmek oldukça önemli olduğundan dolayı T2DM gelişme riski yüksek olan hastaların erken tespitine yönelik çok sayıda araştırma mevcuttur. İlgili çalışmalardan bir kısmı Tablo 2.1’de özetlenmiştir.

**Tablo 2.1.** Diyabetle ilgili bazı veri çalışmaları

| Yazar                       | Yıl  | Veri kaynağı                                | Hasta sayısı | İnceleme konusu                                                           | İnceleme tekniği                                                                              |
|-----------------------------|------|---------------------------------------------|--------------|---------------------------------------------------------------------------|-----------------------------------------------------------------------------------------------|
| (Özge Nur Ergün, 2021)      | 2021 | UCI Makine Öğrenme Deposu                   | 520          | Erken Evre Diyabet Tahmini                                                | KNN, NB, Karar Ağacı, DVM, Rastgele Orman, XGBoost, Yapay Sinir Ağları, Evrimsel Sinir Ağları |
| (Ahmed Kareem Oleiwi, 2020) | 2020 | UCI Makine Öğrenme Deposu                   | 520          | Diyabetin Erken Aşamada Karşılaştırmalı Analizi ve Risk Tahmini           | Rastgele Orman, Radyal Taban Fonksiyon Ağı, Çok Katmanlı Algılayıcılar                        |
| (Smriti Gupta, 2021)        | 2021 | PIMA Indian Diabetes                        | 768          | Diyabetin sınıflandırılmasını analiz etmek ve performans karşılaştırması  | Destek vektör makinesi, Naive Bayes                                                           |
| (Maniruzzaman, 2020)        | 2020 | Ulusal Sağlık ve Beslenme İnceleme Anketi   | 6561         | Diyabet hastalığının tahmini ve risk faktörleri                           | NB, Karar Ağacı, DVM, Rastgele Orman, AdaBoost, Lojistik Regresyon                            |
| (Allaoui, 2022)             | 2022 | Tromso 5 defa kan örneği veren katılımcılar | 234          | Diyabet tahmininde biyobelirteçler                                        | Lojistik regresyon                                                                            |
| (Kibria HB, 2022)           | 2022 | PIMA Indian Diabetes                        | 768          | Diyabet hastalığının tahmininde makine öğrenme tekniklerinin kullanılması | Yapay Sinir Ağları, Rastgele Orman, Destek vektör makinesi, LR, Adaboost, XGBoost             |

|                               |      |                               |     |                                                                                                                              |                                                                                                                   |
|-------------------------------|------|-------------------------------|-----|------------------------------------------------------------------------------------------------------------------------------|-------------------------------------------------------------------------------------------------------------------|
| (Sumbal Malik, 2020)          | 2020 | Frankfurt hastanesi veri seti | --  | Kadınlarda diyabetin erken tanısında makine öğrenme tekniklerinin karşılaştırılması                                          | Rastgele Orman, Karar Ağacı, NB, Adaboost, K en yakın komşu, Lojistik Regresyon, Destek vektör makinesi, BayesNet |
| (Han Wu, 2018)                | 2018 | PIMA Indian Veri Seti         | 768 | Tip 2 diyabetin tahmininde veri madenciliği tekniklerine dayalı yeni bir model incelemesi                                    | K-ortalama ve lojistik regresyon                                                                                  |
| (J. Pradeep Kandhasamy, 2015) | 2015 | UCI Makine Öğrenme Deposu     | 520 | Veri madenciliği teknikleri kullanarak diyabeti tahmin etmek için kullanılan algoritmaların performansının karşılaştırılması | KNN, J48, Destek Vektör Makinesi, Rastgele Orman                                                                  |

Wu ve arkadaşları T2DM'nin artan sıklığı nedeniyle, tip 2 diyabetin tahmininde veri madenciliği tekniklerine dayalı yeni bir model önermişlerdir. PIMA veri setini kullanmışlardır. K-ortalama ve lojistik regresyon algoritmalarını kullanarak yeni bir model önermişlerdir. Önerdikleri modelin diğer araştırmacıların deneysel sonuçlarından daha yüksek tahmin doğruluğuna sahip olduğunu göstermişlerdir (Han Wu, 2018).

Maniruzzaman ve diğerleri, diyabeti sınıflandırmak ve tahmin etmek için bir makine öğrenimi paradigması kullanmışlardır. Diyabet sınıflandırması için dört makine öğrenme algoritması, yani Naïve Bayes, karar ağacı, AdaBoost ve rastgele orman kullanmışlardır. Ulusal Sağlık ve Beslenme İnceleme Anketinden elde edilen 2009–2012 yıllarında yürütülen diyabet veri setinden yararlanmışlar ve 6561 katılımcının verisini kullanmışlardır. Lojistik regresyon, diyabet hastalığı için risk faktörlerini p değeri ve olasılık oranına göre belirlemek için kullanılmış ve model, yaş, eğitim, vücut kitle endeksi, sistolik kan basıncı, diyastolik kan basıncı, direkt kolesterol ve toplam kolesterol gibi 14 faktörün 7'sinin diyabet için risk faktörleri olduğunu göstermiştir. Makine Öğrenme tabanlı sistemin genel doğruluğu %90,62'dir. Lojistik Regresyon tabanlı özellik seçimi ve Rastgele Orman tabanlı sınıflandırıcının kombinasyonu %94,25 doğruluk ile 0,95 AUC vermiş ve bu kombinasyonun diğer makine öğrenme algoritmalarına göre daha iyi performans gösterdiği görülmüştür (Maniruzzaman, 2020).

İslam ve diğerleri, diyabeti erken veya başlangıç aşamasında tahmin etmek için rastgele orman, lojistik regresyon ve Naive Bayes algoritması gibi veri madenciliği tekniklerini kullanmışlardır. Eğitim amacıyla 10 kat çapraz doğrulama ve yüzde bölme teknikleri kullanmışlardır. Anketler yoluyla doğrudan Bangladeş'teki bir hastaneden 529 kişiden diyabetik ve diyabetik olmayan verileri toplamışlardır. Deneysel sonuçlar, rastgele ormanın diğer algoritmalarla kıyasla daha iyi

performans gösterdiğini göstermektedir. Ancak, en gelişmiş karşılaştırma eksiktir ve elde edilen doğruluk açıkça bildirilmemiştir (Islam, 2020).

Ergün ve diğerleri 520 hastaya ait verileri kullanarak, diyabetin erken tanısında kullanmak amacıyla makine öğrenme tekniklerini karşılaştırmışlardır. K-en Yakın Komşu, Destek Vektör Makinesi, Naive Bayes, Karar Ağacı, Rastgele Orman, XGBoost, Yapay Sinir Ağları ve Evrişimsel Sinir Ağları algoritmalarını kullandıkları çalışmada en iyi sonucu Evrişimsel sinir ağı kullanılarak elde etmişlerdir. Modelin doğruluğu %99.04, kesinliği %100, hassasiyet oranı %98.63 ve f skoru %99.31 olarak ölçmüşlerdir. Elde ettikleri sonuçlar doğrultusunda geliştirdikleri algoritmanın diyabetin erken tanısında bir soru seti olarak kullanılabileceğini göstermişlerdir (Özge Nur Ergün, 2021).

Diyabetin erken tanısında makine öğrenme yöntemlerini karşılaştırarak en iyi sınıflandırıcı bulmayı amaçlayan benzer bir çalışmada ise, rastgele orman, çok katmanlı algılayıcı algoritması ve radyal taban fonksiyon ağı kullanmışlardır. Veri seti olarak UCI Makine Öğrenme Deposunu kullanan çalışmada, değişken seçiminde diyabete sebep olabilecek semptomlardan yararlanmışlardır. En yüksek doğruluk oranına sahip sınıflandırmanın %98.80 ile Radyal Taban Fonksiyon Ağı olarak bulmuşlardır (Ahmed Kareem Oleiwi, 2020).

Başka bir çalışmada ise, yapay sinir ağı (YSA), rastgele orman, destek vektör makinesi (DVM), lojistik regresyon (LR), AdaBoost, XGBoost bir ensemble sınıflandırıcı ile birlikte diyabet hastalığını teşhis etmek için kullanılmıştır. En yüksek ROC AUC değeri %95 ile XGBoost bulmuşlardır. (Kibria HB, 2022)

Malik ve diğerleri, kadınlarda diyabeti öngörmek için makine öğrenimi tekniklerinin karşılaştırmalı bir analizini yapmışlardır. Önerdikleri sistem, Almanya'daki bir hastanenin diyabet veri seti üzerine çalışılmış ve makine öğrenme yöntemleri olarak Naive Bayes, Rastgele Orman, AdaBoost, K en yakın komşu, karar ağacı, BayesNet, Lojistik regresyon, Destek Vektör Makinesi uygulamışlardır. Sonuçlar, diğer algoritmalara kıyasla K en yakın komşu, rastgele orman ve karar ağacının üstünlüğünü göstermektedir (Sumbal Malik, 2020).

Kumari ve diğerleri, yaygın olarak kullanılan üç denetimli makine öğrenme algoritmasını bir bütün halinde kullanan yumuşak bilgisayar tabanlı bir diyabet tahmin sistemi sunmuştur. Değerlendirme amacıyla PIMA veri seti kullanmışlardır. Rastgele orman, lojistik regresyon ve Naive Bayes kullanmış ve performanslarını son teknoloji bireysel ve topluluk yaklaşımlarıyla

karşılaştırmışlardır ve sistemleri %79 doğrulukla daha iyi performans göstermiştir (Saloni Kumari, 2021).

Gupta ve diğerleri diyabet sınıflandırması için naive bayes ve destek vektör makine algoritmalarını desteklemişlerdir. PIMA Indian Diabetes veri setini kullanmışlardır. Ayrıca, modelin doğruluğunu artırmak için öznitelik seçimi tabanlı bir yaklaşım ve k-katlı çapraz doğrulama kullanmışlardır. Deneysel sonuçlar, destek vektör makinesinin Naive Bayes modeline göre üstünlüğünü göstermiştir (Smriti Gupta, 2021).

Hussain ve Naaz, diyabet tahmini için 2010-2019 yılları arasında sunulan makine öğrenimi modellerinin kapsamlı bir incelemesini sunmuşlardır. Geleneksel denetimli makine öğrenimi modellerini, doğruluk ve verimlilik açısından sinir ağı tabanlı algoritmalarla karşılaştırmışlardır. Değerlendirme amacıyla Matthews korelasyon katsayısını kullanmışlar ve diğer algoritmalarla kıyasla Naive Bayes ve rastgele ormanın üstünlüğünü gözlemlemişlerdir (Arooj Hussain, 2021).

Başka bir meta-analiz çalışmasında ise, tip 2 diyabet hastalığının makine öğrenme algoritmalarıyla genel tahmin yeteneğini belirlemek amaçlanmıştır. 2010-2021 yılları arasındaki 34 çalışmayı incelemişlerdir. Makine öğrenimi modellerinin özellikle karar ağacı ve yapay sinir ağı modellerinde, T2DM'nin öngörülmesinde yüksek doğruluk verdiği sonucu göstermişlerdir (Micheal O. Olusanya, 2022).

Birkaç klinik çalışma, sağlıklı beslenme ve fiziksel aktiviteye odaklanan yoğun yaşam tarzı değişikliklerinin T2DM gelişimini önleyebileceğini göstermiştir (Reduction in the Incidence of Type 2 Diabetes with Lifestyle Intervention or Metformin, 2002) (Mackerras, 2003). Bu yüzden, T2DM geliştirme risklerini azaltmak için bu riske sahip bireyleri belirlemek önem arz etmektedir. T2DM'nin etiyolojisi, genetik ve çevresel faktörler arasındaki karmaşık ve yeterince anlaşılmayan etkileşimleri içerir. Çalışmalar vücut kitle indeksi, ailede DM öyküsü, açlık kan şekeri ve glikozillenmiş hemoglobinin (HbA1c), T2DM geliştirme riski daha yüksek olan kişileri belirlemek için yaygın olarak kullanılan klinik risk faktörleri olduğunu göstermiştir (Meigs, 2009).

Risk skoru modelleriyle ilgili çeşitli çalışmalarda ise, temel klinik bilgilere kan biyobelirteçlerinin eklenmesinin, özellikle glisemik süreçler, ürik asit ve lipidlerle ilgili biyobelirteçler olmak üzere, tahmin yeteneğini geliştirdiğini göstermiştir (Wilson PW, 2007).

75 yaş üstü yaşlı bir Çinli popülasyonda yeni başlangıçlı T2DM'nin bir göstergesi olarak trigliseridin yüksek yoğunluklu lipoprotein kolesterol (TG/HDL-C) oranına rolünü değerlendirmeyi amaçlayan bir çalışmada, TG/HDL-C oranı, yaşlı erişkinlerde T2DM'yi öngörmeye basit ama etkili bir gösterge olduğu bulunmuştur (Hongzhou Liu, 2022).

Tip 2 diyabetli hastalarda çeşitli biyokimyasal ve hematolojik parametreleri araştıran bir çalışmanın sonuçlarına göre kontrol grubuna göre T2DM grubunda SGPT, alkalın fosfataz, üre, serum kreatinin, toplam kolesterol, trigliserit ve LDL ortalama değerlerinde bir artış göstermiş ve T2DM grubunda HDL ortalama değeri kontrol grubuna göre önemli ölçüde azalmıştır. Ortalama hemoglobin, RBC, MCV, MCHC ve MCH değerleri T2DM grubunda kontrol grubuna göre anlamlı derecede azalırken, kırmızı hücre dağılım genişliği, kontrol grubuna kıyasla T2DM grubunda önemli ölçüde arttığı görülmüştür (Alshehri, 2017).

### **2.3.5. Makine Öğrenmesi ile Yapılan Literatür Çalışmaları**

Makine öğrenimi (ML), kalıpları ortaya çıkarmak ve davranışları tahmin etmek için veriye dayalı teknikleri kullanan bilgisayar biliminde yapay zekanın bir alt alanıdır. Son yıllarda, hastalık teşhisi ve tedavisinde doğru ve etkili olduğu kanıtlanmış olan makine öğrenimi teknikleri tıp ve sağlık alanında yaygın olarak uygulanmaktadır (Topol E. J., 2019). Makine öğrenimi algoritmalarına dayalı tahmine dayalı modeller, diyabetli hastaların belirlenmesinde faydalı olabilmekte ve diyabet risk faktörlerinde gözden kaçabilecek gizli kalıpları keşfetmeye yardımcı olabilmektedir (Lai H H. H., 2019 )

Klinik değerlendirmelerden yüksek kaliteli sonuçlar elde etmek için uygulanan makine öğrenimi yaklaşımlarını ve karşılaştırmalı performanslarını özetleyen bir çalışmada, 3918 çalışma incelenmiş ve sonuçlar, makine öğreniminin %86'lık yüksek bir hassasiyeti korurken %95'lik bir hassasiyete ulaşabileceğini göstermiştir (Abdelkader, 2021). Lojistik regresyon, destek vektör makineleri, yapay sinir ağı, k-en yakın komşular, karar ağacı, rastgele orman ve aşırı gradyan artırma (XGBoost) gibi T2DM'nin tahmini için çok sayıda makine öğrenme algoritması kullanılmıştır (Nongyao Nai-arun, 2019).

2013-2019 yılları arasında diyabeti tahmin etmek için yapılan çalışmalarda da, her bir makine öğrenimi sınıflandırıcısının kullanım sıklığı hesaplanmış, Yapay Sinir Ağı, Destek Vektör Makinesi, Karar Ağacı ve Naive Bayes, diyabet tahmini için en çok kullanılan sınıflandırıcılar olduğu gösterilmiştir. Aynı zamanda bu yıllar arasında çok az kullanılan makine öğrenimi sınıflandırıcılarını da incelemişlerdir ve bunlar REPTree, M5P, KStar, OneR, PART, ZeroR,

SMO, BayesNettir. En yüksek doğruluk, REPTree tarafından elde edilen %74.48'dir (Larabi Marie-Sainte, 2019).

Sneha et al. sınıflandırma yöntemlerinin doğruluğunu artırmak için optimal öznelik seçim yöntemini kullanmış ve Naive Bayes yönteminin en iyi doğruluğu verirken rastgele orman en yüksek özgüllüğü verdiğini göstermiştir (Sneha, 2019).

Yakın zamanda yapılan bir meta-analiz, makine öğrenimi modellerinin topluluk ortamlarında T2DM'yi tahmin etmek için iyi bir ayırım yeteneğini doğrulamış ve yapay sinir ağının en iyi performansı gösterdiğini, ardından lojistik regresyon, karar ağaçları ve rastgele ormanların geldiğini göstermiştir (Kushan De Silva, 2020).

Kandhasamy ve Balamurali, birden çok sınıflandırıcı DVM, J48, K-Nearest Neighbors ve Random Forest kullanmışlardır. Sınıflandırma, UCI deposundan alınan bir veri seti üzerinde gerçekleştirilmiştir. Sınıflandırıcıların sonuçları doğruluk, duyarlılık ve özgüllük değerlerine dayalı olarak karşılaştırılmıştır. Ön işleme yapılmadan karar ağacı J48 sınıflandırıcısının %73,82 ile en yüksek doğruluk oranına sahip olduğunu, KNN (k=1) ve Random Forest sınıflandırıcılarının ise verileri ön işleme tabi tuttukten sonra %100 ile en yüksek doğruluk oranına sahip olduğunu bildirmişlerdir (J. Pradeep Kandhasamy, 2015).

Ayrıca Yuvaraj ve Sripreethaa, Random Forest, Decision Tree ve Naïve Bayes dahil olmak üzere üç farklı ML algoritması kullanarak diyabet tahmini için bir uygulama sunmuştur. Pima Indian Diabetes veri seti, ön işleme tabi tutulduktan sonra kullanılmışlardır. Rastgele orman algoritmasının %94 ile en yüksek doğruluk oranına sahip olduğunu göstermişlerdir (Yuvaraj, 2019).

Ayrıca, Tafa ve ark. diyabeti tahmin etmek için yeni bir entegre geliştirilmiş DVM ve Naive Bayes modeli önermiştir. Önerilen birleşik algoritmalar, tahminin doğruluğunu %97.6'ya ulaşacak şekilde geliştirmiştir. Bu değeri, sırasıyla %95.52 ve %94.52'ye ulaşan DVM ve Naive Bayes'in performansı ile karşılaştırmışlardır (Zhilbert Tafa, 2015) .

### **2.3.6. Diyabette Kan Hücreleriyle Yapılan Çalışmalar**

Diyabetli kişilerde sıklıkla hematolojik anormallikler vardır. Bunlar anemi ve diğer kırmızı hücre problemlerini içerir. Beyaz kan hücresi ve trombosit anormallikleri de diyabetli kişilerde yaygındır. Çoğu hücre tipini etkileyen hematolojik anormallikler diyabetli kişilerde yaygındır.

WBC deęişiklikleri, diyabetin bařlangıcından önce veya diyabetin bařlangıcında iltihaplanma veya immünolojik sorunları vurgulayabilmektedir. Nötrofil arızası, diyabetliler arasında daha fazla enfeksiyona ve enfeksiyonların tedavi edilmesindeki zorluklara katkıda bulunmaktadır. Trombosit anormallikleri, diyabette kardiyovasküler hastalığa katkıda bulunmaktadır. Kan řekeri kontrolünün iyileřtirilmesini, WBC ve trombosit fonksiyonunu iyileřtirebileceęini görmüşlerdir (Adane T, 2021).

Schmidt ve arkadaşları Tip 2 diabetes mellitus ve kardiyovasküler hastalıkların ortak öncülleri olduęunu ileri sürmüşlerdir. Enflamasyon belirteçleri koroner kalp hastalığını öngördüęünden ve Tip 2 diyabetli hastalarda büyütüldüęünde, insanlarda tip 2 diyabet geliřtirip geliřtirmeyeceklerini tahmin edip etmediklerini arařtıran bir çalışmaları, 7 yıl boyunca hastaları takip etmişlerdir. Vücut kitle indeksi ve bel-kalça oranı için yapılan ayarlamalar ilişkileri önemli ölçüde azaltmış olsa da, beyaz hücre sayısının en yüksek dörtte birindeki bireylerin diyabet geliřtirme olasılığı %50 daha yüksek olduęu bulmuşlardır (Schmidt MI, 1999).

En az 3 aylık süre boyunca insülin tedavisi almasına raęmen uygun olmayan kan řekeri yönetimi (HbA1c deęeri > %7) olan 135 T2DM ve 121 saęlıklı kiři dahil eden bir çalışmada, bulgular WBC, NEU, LYM ve MONO sayıları DM grubunda daha yüksek olduęunu göstermiştir (Muhammed Kizilgul, 2018).

Tip 2 diyabetli hastalarda tam kan sayımını diyabetik olmayan saęlıklı kontrollerle karřılařtıran bir incelemede, RDW deęerinin diyabetik hastalarda saęlıklı bireylere göre anlamlı derecede yüksek olduęunu ve özellikle kontrolsüz glisemide daha yüksek olduęunu görmüşlerdir. (Gkrania-Klotsas E, 2010 ).

T2DM olan ve olmayan bireylerde 30 yıl boyunca kan deęişkenlerinin konsantrasyonlarındaki tanı öncesi ve sonrası deęişiklikleri tanımlamayı ve tanı öncesi kan deęişkenlerinin öngörücü potansiyelini belirlemeyi amaçlayan çalışmada, 1986 ve 2016 arasında beř zaman noktasında kan örneęi veren 234 katılımcı arasından 130 katılımcı T2DM geliřtirmemiş ve kontrol olarak kullanılmıştır; 104 katılımcı ise üçüncü zaman noktasından sonra T2DM'yi geliřtirmiş ve vaka olarak dahil edilmiştir. Olgular ve kontroller lipidler, serbest T3, HbA1c, glukoz ve GGT'de farklı boylamsal deęişiklikler yaşamıştır.

Sonuçlara göre BMI, fiziksel aktivite, yüksek kan basıncı ve T2DM aile öyküsü ile kombinasyon halinde seçilen kan değişkenleri (lipidler, HbA1c, GGT), T2DM teşhisinden 1-7 yıl önce mükemmel tahmin yeteneğine ve teşhisten 15 yıl öncesine kadar kabul edilebilir tahmin yeteneğine sahip olduğu bulunmuştur (Allaoui G, 2022).

Suudi Arabistan'da tip 2 diabetes mellituslu (T2DM) hastalarda çeşitli biyokimyasal ve hematolojik parametreleri araştırmak ve diyabetik olmayan deneklerle (kontrol grubu) karşılaştırmayı amaçlayan benzer bir çalışmada, sonuçlar kontrol grubuna göre T2DM grubunda SGPT, alkalın fosfataz, üre, serum kreatinin, toplam kolesterol, trigliserit ve LDL ortalama değerlerinde bir artış göstermiş olduğunu gözlemlemişlerdir. Ortalama hemoglobin, RBC, MCV, MCHC ve MCH değerleri T2DM grubunda kontrol grubuna göre anlamlı derecede azalmıştır. Buna karşılık, kırmızı hücre dağılım genişliği, kontrol grubuna kıyasla T2DM grubunda önemli ölçüde arttığı gözlemlemişlerdir (Alshehri, 2017).

WBC sayısının obez olmayan bireyler arasında T2DM olayının tahmini için bir gösterge olarak kullanılıp kullanılamayacağını belirlemeyi amaçlayan başka bir çalışmada, 10 yıl boyunca gözlemlenen bireylerin verileri sonucunda, daha yüksek bir WBC sayısı, toplulukta yaşayan obez olmayan Koreli yetişkinler arasında HOMA-IR dahil olmak üzere diğer ilişkili değişkenlerden bağımsız olan gelecekteki T2DM olayını tahmin ettiği görmüşlerdir. Klinik uygulamada, yaygın olarak ölçülen WBC sayısı, T2DM'ye duyarlı obez olmayan bireylerin tahminini kolaylaştırdığı sonucuna varmışlardır (Park JM, 2021).

Sistematik rastgele örnekleme tekniği ile seçilen 268 (134 tip 2 diyabet hastası ve 134 kontrol) çalışma katılımcısı üzerinde yapılan karşılaştırmalı bir kesitsel mevcut çalışmada ise, tip 2 diyabetli hastalarda toplam beyaz kan hücresi sayısının, mutlak nötrofil, lenfosit, eozinofil ve bazofil sayımlarının, kırmızı kan hücresi dağılım genişliğinin, trombosit sayısının ve ortalama trombosit hacminin kontrol grubuna kıyasla anlamlı olarak daha yüksek olduğunu bulmuşlardır. Öte yandan, tip 2 diyabetli hastalarda ortalama hemoglobin kontrol grubuna göre anlamlı derecede düşük olduğunu görmüşlerdir (Mesay Arkew, 2021).

### 3. GEREÇ ve YÖNTEM

#### 3.1. Veri toplama

Bu çalışmada, Akdeniz Üniversitesi Tıp Fakültesi Klinik Araştırma Etik kurulunun 2012-KAEK-20 kodlu kararı alınmış olup Tip 2 diyabete sahip bireylerin ve kontrol grubu hastalarının verileri toplanmıştır. Uzman hekimler tarafından Tip 2 Diyabet tanısının konulduğu hastalara ait ilk kan testi sonucu ve hiç diyabet tanısı konulmamış kontrol hastalarına ait veriler toplanmıştır. Veri seti Akdeniz, Marmara ve Ege coğrafi bölgelerinde bulunan özel hastanelerden toplanan, toplam 126 hastanın (66 diyabetik olmayan vaka ve 60 diyabetik vaka) bilgilerinden oluşmaktadır. Hastalara ait veriler toplanırken kişilerin ad, soyad, tc kimlik numarası gibi kişisel verileri alınmamıştır.

Tablo 3.1'de gösterildiği gibi değişken adları ve veri tipleri ile birlikte 44 biyolojik değişken vardır. Değişkenler doğada kategorik ve sayısaldir. Diyabet hastalarının ilk tanısının konulduğu tarih olan tanı tarihi, hasta verilerinin alındığı tarih, hastanın yaşı, cinsiyeti, doğum yeri, ikamet ettiği şehir, kandaki lökosit sayısını veren beyaz kan hücre sayısı WBC, şeker hastalığının tanısını koymada sıklıkla kullanılan glikozillenmiş hemoglobin HBA1C, kırmızı kan hücresi veya eritrosit olarak bilinen RBC, hemoglobin HGB, kırmızı kan hücrelerinin hacminin, dolaşımdaki kanın hacmine oranını ifade eden Hematokrit yani HCT, eritrosit olarak bilinen kırmızı kan hücrelerinin boyutu hakkında bilinen MCV, ortalama eritrosit hemoglobini MCH, ortalama korpüsküler hemoglobin konsantrasyonu MCHC, vücut damarlarında kanın pıhtılaşmasını sağlayan hücre pulcukları PLT, kandaki trombositlerin büyüklüğünün dağılımını veren PDW, ortalama trombosit hacmi MPV, vücutta bir enfeksiyon olup olmadığını gösteren PCT, bağışıklık sistemine saldıran bakterilere karşı savaşan beyaz kan hücreleri yani nötrofil NEUT, kanın bileşiminde yer alan beyaz renkli hücreler yani lenfositler LYM, viral ve bakteriyel enfeksiyonlara karşı direnç gösteren ve koruyan Monositler MONO, eozinofil yani bağışıklık sistemi konusunda insan vücudunda önemli bir rolü olan EO, bağışıklık sistemini ve insan vücudunu koruyan bir beyaz kan hücresi BASO, kandaki potasyum K değeri, kandaki sodyum Na değeri, kalsiyum değeri CA, tiroit uyarıcı hormon TSH değeri, kolesterol bilgisi, iyi kolesterol seviyesi HDL, kötü kolesterol seviyesi LDL, sistolik kan basıncı, diastolik kan basıncı, glukoz açlık, trigliserid, üre, ürik asit, CRP, B12 ve kreatinindir. Bu çalışmada kullanılan veriler, kişisel sağlık verilerinden elde edilen sonuçlardır.

Bu çalışmada kullanılan veri setinin temel amacı, makine öğrenmesi teknikleri kullanılarak hastalara ait tam kan sayımı sonuçlarından tip 2 diyabet hastalığının önceden öngörülmesi

konusunun ele alınmasıdır. Veri setinde sağlanan seçilmiş özelliklere dayanarak, olası hastaları ve hastalığa olan eğilimlerini tahmin etmek için çalışma genişletilebilir.

**Tablo 3. 1** Veri Setinde Kullanılan Özellikler ve Tipleri

| S. No | Özellik Adı           | Tipi         |
|-------|-----------------------|--------------|
| 1     | Tanı Tarihi           | DateTime     |
| 2     | Tarih                 | DateTime     |
| 3     | Yaş                   | Int          |
| 4     | Cinsiyet              | VarChar(1)   |
| 5     | Doğum Yeri            | VarChar(100) |
| 6     | Yaşadığı Şehir        | VarChar(100) |
| 7     | HbA1C                 | Decimal      |
| 8     | Kolestrol             | Decimal      |
| 9     | HDL Kolestrol         | Decimal      |
| 10    | LDL Kolestrol         | Decimal      |
| 11    | Trigliserid           | Decimal      |
| 12    | Sistolik Kan Basıncı  | Decimal      |
| 13    | Diastolik Kan Basıncı | Decimal      |
| 14    | Glukoz Açlık          | Int          |
| 15    | Üre                   | Decimal      |
| 16    | Ürik Asit             | Decimal      |
| 17    | Kreatinin             | Decimal      |
| 18    | CRP                   | Decimal      |
| 19    | B12                   | Int          |
| 20    | WBC                   | Decimal      |
| 21    | RBC                   | Decimal      |
| 22    | HGB                   | Decimal      |
| 23    | HCT                   | Decimal      |
| 24    | MCV                   | Decimal      |
| 25    | MCH                   | Decimal      |
| 26    | MCHC                  | Decimal      |
| 27    | PLT                   | Decimal      |
| 28    | PDW                   | Decimal      |
| 29    | MPV                   | Decimal      |
| 30    | PCT                   | Decimal      |
| 31    | NEUT                  | Decimal      |
| 32    | NEUT%                 | Decimal      |
| 33    | LYM                   | Decimal      |
| 34    | LYM%                  | Decimal      |
| 35    | MONO                  | Decimal      |
| 36    | MONO%                 | Decimal      |
| 37    | EO                    | Decimal      |
| 38    | EO%                   | Decimal      |
| 39    | BASO                  | Decimal      |
| 40    | BASO%                 | Decimal      |
| 41    | K                     | Decimal      |
| 42    | NA                    | Decimal      |
| 43    | CA                    | Decimal      |
| 44    | TSH                   | Decimal      |

Tarih aralığı olarak 01.01.2016 ve 01.05.2022 arasındaki son 6 seneye ait veriler seçilmiştir. Hastalara ait veriler bir HBYS firması olan Medisoft programını kullanan hastanelerin veri tabanlarından toplanılmıştır. Veri tabanı işlemleri için ücretsiz bir yazılım olan MySQL kullanılmıştır. Stored prosedure yardımıyla veriler çekilmiştir. Stored procedure yazımında ICD (International Classification of Diseases) kodlarından yararlanılmıştır. T2DM tahmini için modeller oluşturulup dağıtılmadan önce virgülle ayrılmış değerler (CSV) dosya formatında hazırlanmıştır. Verilerin hazırlanacağı arayüz Pascal dilinde yazılarak, Delphi programı ile gerçekleştirilmiştir. Şekil 3.1 verilerin hazırlandığı arayüz gösterilmiştir. Database tarafında

oluşturulan tabloya HKIMLIKID kolonu eklenmiş olup, bir hastaya ait birden fazla kan testi sonucu var ise aynı ID ile gösterilip, karşılaştırma için kolay erişimi sağlanmıştır.

| Kimlik Bilgileri |            |            |          |
|------------------|------------|------------|----------|
| HKIMLIKID        | 0          | Yaş        | 69       |
| Cinsiyet         | K          | Doğum Yeri | ACIPINAR |
| İKAMETGAH YERİ   | İSTANBUL   | AKTİF      | Evet     |
| TARİH            | 11.02.2020 |            |          |

| Tam Kan Sayımı (Hemogram) |      |                 |      |
|---------------------------|------|-----------------|------|
| WBC                       | 6,98 | PLT (TROMBOSİT) | 0    |
| RBC                       | 4,56 | MPV             | 9,1  |
| HGB                       | 12,2 | PCT             | 0,26 |
| HCT                       | 37,3 | NEUT            | 3,68 |
| MCH                       | 26,8 | NEUT%           | 52,8 |
| MCHC                      | 32,7 | LYMPH           | 0    |
| MCV                       | 81,8 | LYMPH%          | 0    |
| LYM                       | 2,6  | RBC             | 4,56 |
| LYM#                      | 0    | RDW             | 13,7 |
| LYM%                      | 2,6  | MID             |      |
| GRAN                      | 0    | MID%            |      |
| GRAN%                     | 0    |                 |      |

| Biyokimya     |      |             |   |
|---------------|------|-------------|---|
| HBA1C         | 6,5  | HBA1C(NGSP) |   |
| HDL KOLESTROL | 100  | HBA1C(IFCC) |   |
| LDL KOLESTROL | 100  | ÜRE         |   |
| GLUKOZ(Açlık) | 100  | ÜRİK ASİT   |   |
| TRİGLİSERİD   | 100  | CRP         |   |
| KOLESTROL     | 171  | RF          | 0 |
| KREATİNİN     | 100  |             |   |
| KAL SİYUM     | 1001 |             |   |
| ALT (SGPT)    | 0    |             |   |
| AST (SGOT)    | 0    |             |   |
| SODYUM        |      |             |   |
| POTASYUM      |      |             |   |
| ALP           | 0    |             |   |

| Hormon    |  |
|-----------|--|
| V B12     |  |
| FERRİTİN  |  |
| TSH       |  |
| Free T4   |  |
| Free T3   |  |
| Troponin  |  |
| Total PSA |  |

Şekil 3. 1 Verilerin Toplandığı Arayüz

### 3.2. Makine öğrenimi araç seti

Deneyisel çalışma, çeşitli istatistiksel ve makine öğrenimi performans değerlendirme ölçümleri için programlama dili R ile birlikte bir IDE (Entegre Geliştirme Ortamı) olarak R Studio kullanılarak gerçekleştirilmiştir. R Studio, çizim, hataları bulma ve çözme ve çalışma alanı yönetimi için araçlar da içerir (<https://www.rstudio.com/products/rstudio/>, 2022). T2DM hastalığının tahmini için keşifsel veri analizi ve model oluşturma süreci için ön koşullar makine öğrenimi ve üçüncü taraf kitaplıkları içe aktarıldı. T2DM tahmini gibi çeşitli sağlık sorunları için Yapay Zeka, Makine Öğrenimi, Derin Öğrenme modellerinin geliştirilebildiği bu açık kaynaklı yazılımda akıllı ve zengin bir paket kütüphanesi bulunmaktadır (Shahid Mohammad Ganie, 2020).

### 3.3. Veri kümesi açıklaması

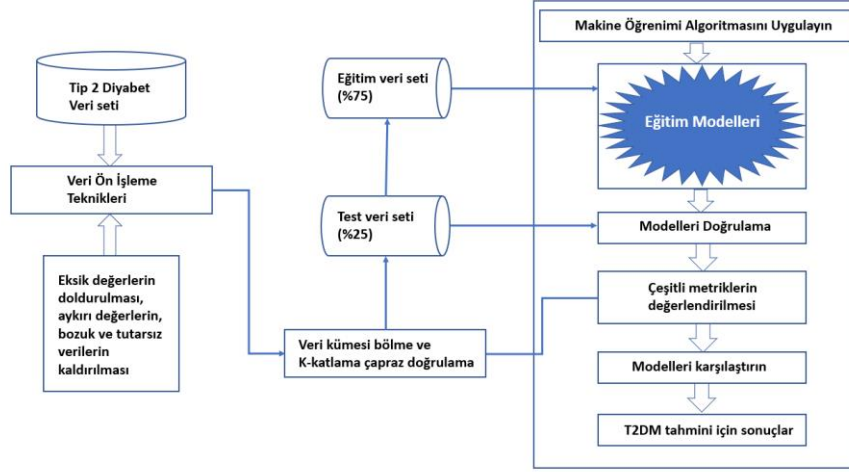
DataFrame tarif() yöntemi, veri kümesinin parametrelerini tanımlamak için kullanılır. Sayım, ortalama, standart, min, maks ve DataFrame ile ilgili diğer sayısal değerler Tablo 3.2'de hesaplanması gösterilmiştir.

**Tablo 3. 2** Veri Kümesinin İstatistiksel Ölçümleri

| Parametre          | Toplam | Min   | Medyan | Ortalama | Maks   |
|--------------------|--------|-------|--------|----------|--------|
| WBC                | 126    | 4.310 | 8.035  | 8.409    | 17.750 |
| RBC                | 126    | 3.000 | 5.020  | 4.976    | 7.280  |
| YAŞ                | 126    | 19.00 | 40.0   | 41.1     | 63.0   |
| HGB                | 126    | 7.40  | 14.29  | 14.12    | 17.70  |
| HCT                | 126    | 27.70 | 42.60  | 42.14    | 54.67  |
| MCV                | 126    | 61.22 | 86.05  | 84.73    | 106.70 |
| MCH                | 126    | 17.08 | 28.90  | 28.42    | 48.00  |
| MCHC               | 126    | 26.70 | 33.50  | 33.47    | 45.00  |
| PLT                | 124    | 101.0 | 262.3  | 261.9    | 441.1  |
| PDW                | 112    | 8.700 | 12.80  | 14.26    | 22.50  |
| MPV                | 115    | 5.630 | 9.70   | 9.42     | 13.40  |
| PCT                | 124    | 0.090 | 0.24   | 0.24     | 0.44   |
| NEUT               | 126    | 0.000 | 4.30   | 5.05     | 15.95  |
| NEUT.              | 126    | 35.69 | 57.45  | 60.48    | 90.60  |
| LYM                | 126    | 0.000 | 2.295  | 2.57     | 37.90  |
| LYM.               | 126    | 0.000 | 30.25  | 27.73    | 51.27  |
| MONO               | 126    | 0.000 | 0.586  | 0.575    | 1.520  |
| MONO.              | 126    | 1.800 | 7.070  | 7.331    | 17.10  |
| EO                 | 126    | 0.000 | 0.120  | 0.182    | 5.000  |
| EO.                | 124    | 0.054 | 1.657  | 2.090    | 9.600  |
| BASO               | 126    | 0.000 | 0.200  | 0.033    | 0.200  |
| BASO.              | 126    | 0.100 | 0.400  | 0.557    | 2.500  |
| KOLESTROL          | 52     | 31.0  | 181.0  | 202.7    | 358.0  |
| HDLKOLESTROL       | 40     | 25.00 | 45.85  | 50.77    | 97.30  |
| LDLKOLESTROL       | 41     | 39.00 | 109.0  | 118.9    | 260.00 |
| GLUKOZACLİK        | 85     | 81    | 109    | 137      | 444    |
| UREA               | 44     | 17.12 | 23.77  | 26.08    | 48.00  |
| URIC ACID          | 47     | 3.300 | 5.300  | 5.576    | 9.500  |
| HBA1C              | 26     | 5.080 | 7.200  | 8.394    | 2.810  |
| TRIGLİSERİD        | 70     | 48.0  | 175.5  | 235.7    | 936    |
| SİSTOLİKANBASINCI  | 14     | 120   | 122.5  | 123.8    | 130.0  |
| DIASTOLİKANBASINCI | 14     | 70.0  | 75.0   | 77.5     | 90.0   |
| KREATİNİN          | 38     | 0.50  | 0.77   | 0.78     | 1.05   |
| K                  | 44     | 3.60  | 4.40   | 4.47     | 5.20   |
| NA                 | 32     | 107   | 140    | 138      | 148    |
| CRP                | 34     | 0.07  | 1.45   | 2.30     | 9.50   |
| CA                 | 20     | 8.60  | 9.40   | 9.41     | 10.10  |
| B12                | 40     | 182   | 281    | 348      | 769    |
| TSH                | 58     | 0.39  | 1.49   | 1.56     | 3.72   |

## Önerilen sistem

Önerilen çerçeve Şekil 3.2'de gösterilmektedir. Veri toplamadan istenen deneysel sonuçlara kadar araştırma sürecimizin genel çalışma mekanizmasını sunar. Kullanılan çerçeve, T2DM'nin daha iyi tahmin edilmesi için makine öğrenimi araçlarına ve tekniklerine dayanmaktadır.



Şekil 3. 2 Araştırma Süreci Önerilen Çerçeve

### 3.4. Veri ön işleme

Bu adımda, toplanan veriler Sql komutlarıyla çekilerek excele aktarılır. Ardından virgülle ayrılmış değerler (CSV) dosya formatına dönüştürülür. Ön işleme, istenen sonuçları elde etmek ve çeşitli makine öğrenme tekniklerinin sınıflandırma kabiliyetini elde etmek ve ham verileri değiştirmek için hayati bir rol oynar (Jazayeri A, 2020). Toplanan verilerde çok fazla eksik değer nedeniyle üre, ürik asit, kreatinin, CRP, B12, sistolik ve diastolik kan basıncı, trigliserid, HDL Kolesterol, LDL Kolesterol, Kolesterol, K, NA, CA, TSH, HBA1C, Glukoz Açlık değişkenleri çıkarılmıştır. Tanı tarihi, doğum yeri ve yaşadığı şehir değişkenleri ise istatistiksel hesaplamalarda kullanılmayacağı için çıkarılmıştır. Geriye kalan değişkenlerin veri kalitesi değerlendirmesini iyileştirmek için ortalama, mod ve medyan gibi değerlendirme yöntemiyle eksik değerleri doldurulmuştur. Kullanılmış olan 24 değişkene ait veriler bulgular Tablo 4.1’de gösterilmiştir.

#### 4. BULGULAR

Makine öğrenmesi için hazırlanan ve 01.01.2016 ve 01.05.2022 tarihleri arasında hastanelerden toplanan kayıtlarından elde edilmiş 126 kişinin verisine ait 24 ayrı niteliğin R Studio istatistiksel analiz yazılımı ile analizi yapılmıştır.

**Tablo 4.1** Kullanılan Veri Kümesinin İstatistiksel Ölçümleri

| Parametre | Toplam | Min   | Median | Mean  | Max    |
|-----------|--------|-------|--------|-------|--------|
| Grup      | 126    |       |        |       |        |
| Cinsiyet  | 126    |       |        |       |        |
| WBC       | 126    | 4.310 | 8.035  | 8.409 | 17.750 |
| RBC       | 126    | 3.000 | 5.020  | 4.976 | 7.280  |
| YAŞ       | 126    | 19.00 | 40.0   | 41.1  | 63.0   |
| HGB       | 126    | 7.40  | 14.29  | 14.12 | 17.70  |
| HCT       | 126    | 27.70 | 42.60  | 42.14 | 54.67  |
| MCV       | 126    | 61.22 | 86.05  | 84.73 | 106.70 |
| MCH       | 126    | 17.08 | 28.90  | 28.42 | 48.00  |
| MCHC      | 126    | 26.70 | 33.50  | 33.47 | 45.00  |
| PLT       | 124    | 101.0 | 262.3  | 261.9 | 441.1  |
| PDW       | 112    | 8.700 | 12.80  | 14.26 | 22.50  |
| MPV       | 115    | 5.630 | 9.70   | 9.42  | 13.40  |
| PCT       | 124    | 0.090 | 0.24   | 0.24  | 0.44   |
| NEUT      | 126    | 0.000 | 4.30   | 5.05  | 15.95  |
| NEUT.     | 126    | 35.69 | 57.45  | 60.48 | 90.60  |
| LYM       | 126    | 0.000 | 2.295  | 2.57  | 37.90  |
| LYM.      | 126    | 0.000 | 30.25  | 27.73 | 51.27  |
| MONO      | 126    | 0.000 | 0.586  | 0.575 | 1.520  |
| MONO.     | 126    | 1.800 | 7.070  | 7.331 | 17.10  |
| EO        | 126    | 0.000 | 0.120  | 0.182 | 5.000  |
| EO.       | 124    | 0.054 | 1.657  | 2.090 | 9.600  |
| BASO      | 126    | 0.000 | 0.200  | 0.033 | 0.200  |
| BASO.     | 126    | 0.100 | 0.400  | 0.557 | 2.500  |

Her bir niteliğin, 1) Diyabetli ve 2) Kontrol gruplarına göre Odds oranı ve P değerleri hesaplanmış Tablo 4.2’de özetlenmiştir. Gruplar birbirinden bağımsızdır. Kategorik değişkenler için ki-kare testi yapılırken, nümerik değişkenler için Mann Whitney U testi uygulanmıştır. Bu analizde diyabetli grubun kontrol grubuna göre etki eden değişkenler ve odds oranları şöyledir; HGB(0,41), BASO (0,78), MCHC (0,69), MONO (0,31), NEUT (0,23), WBC (0,12), RBC (0,13), EO (0,27), HCT (0,17). Analiz çalışmalarında anlamlılık değeri  $p < 0.05$  alınmıştır.

**Tablo 4.2** Tek deęişkenli analiz çalışması sonucu özet tablosu

| Parametre | Diyabetli Grup | Kontrol Grubu | Odds Oranı | Güven Aralığı | p      |
|-----------|----------------|---------------|------------|---------------|--------|
| WBC       | 8,55           | 8,27          | 0,99       | (0,95-1,02)   | 0,567  |
| RBC       | 5,01           | 4,93          | 1,02       | (1,01-1,03)   | 0,458  |
| YAŞ       | 37             | 44            | 1,02       | (1,01-1,03)   | <0,001 |
| HGB       | 14,54          | 13,73         | 0,94       | (0,90-0,99)   | 0,021  |
| HCT       | 42,63          | 41,70         | 0,99       | (0,97-1)      | 0,316  |
| MCV       | 85,20          | 84,30         | 0,99       | (0,98-1)      | 0,476  |
| MCH       | 29,15          | 27,75         | 0,97       | (0,94-0,99)   | 0,022  |
| MCHC      | 34,14          | 32,85         | 0,91       | (0,88-0,95)   | <0,001 |
| PLT       | 259,47         | 264,03        | 1,00       | (0,99-1)      | 0,765  |
| PDW       | 14,37          | 14,16         | 0,99       | (0,97-1,02)   | 0,523  |
| MPV       | 9,50           | 9,35          | 0,97       | (0,91-1,04)   | 0,539  |
| PCT       | 0,24           | 0,24          | 1,11       | (0,30-4,12)   | 0,790  |
| NEUT      | 5,38           | 4,74          | 0,97       | (0,94-1,01)   | 0,198  |
| NEUT.     | 60,36          | 60,58         | 1,00       | (0,99-1)      | 0,913  |
| LYM       | 2,37           | 2,76          | 1,00       | (0,98-1,03)   | 0,326  |
| LYM.      | 28,24          | 27,25         | 0,99       | (0,99-1)      | 0,641  |
| MONO      | 0,61           | 0,54          | 0,71       | (0,49-1,04)   | 0,082  |
| MONO.     | 7,43           | 7,23          | 0,99       | (0,95-1,02)   | 0,641  |
| EO        | 0,24           | 0,12          | 0,86       | (0,70-1,04)   | <0,001 |
| EO.       | 2,00           | 2,17          | 1,01       | (0,96-1,07)   | 0,537  |
| BASO      | 0,04           | 0,02          | 0,99       | (0,98-1)      | <0,001 |
| BASO.     | 0,58           | 0,53          | 0,94       | (0,77-1,13)   | 0,223  |

#### 4.1. Lojistik Regresyon

Aynı veri setindeki 24 deęişkenin gruplara göre sonucu nasıl etkilediğini ortaya koymak amacıyla yapılan çok deęişkenli analiz sonuçları aşağıdaki Tablo 4.3’de özetlenmiştir. Güven aralığı % 95 alınarak elde edilen anlamlı sonuçlar ise şöyledir; Yaş (1,16 (1,08-1,27)), Cinsiyet (2,26 (3,63 – 1,55)), WBC (4,84 (1,63 – 27,6)), RBC (0,54 (0,74 – 0,43)), HGB (2,46 (1,57 – 8,18)), HCT (0,33 (< ,001– 2,28)), MCV (1,66 (0,18 – 16,4)), MCH (0,23 (< ,001- 1,38)), MCHC (0,002 (< ,0001– 0,7)), PLT (0,95 (0,89 - 1,01)), PDW (0,80 (6,12 - 1,03)), MPV (3,69 (6,85 – 1,97)), PCT (0,99 (1,38 - 0,92)), NEUT (1,27 (0,04 – 0,94)), NEUT. (1,53 (1,22 – 2,10)), LYM (1,33 (1,07 - 1,69)), LYM. (1,24 (0,08 - 1,52)), MONO (0,007 (< ,0001– 0,2)), MONO. (2,47 (1,10 – 3,78)), EO (0,001 (0,0001 -0,001)), EO. (5,43 (2,32 - 24,7)), BASO (0,003 (0,0004 – 0,02)), BASO. (4,06 (0,32 – 8,34)).

**Tablo 4.3** Çok deęişkenli analiz çalışması sonucu özet tablosu

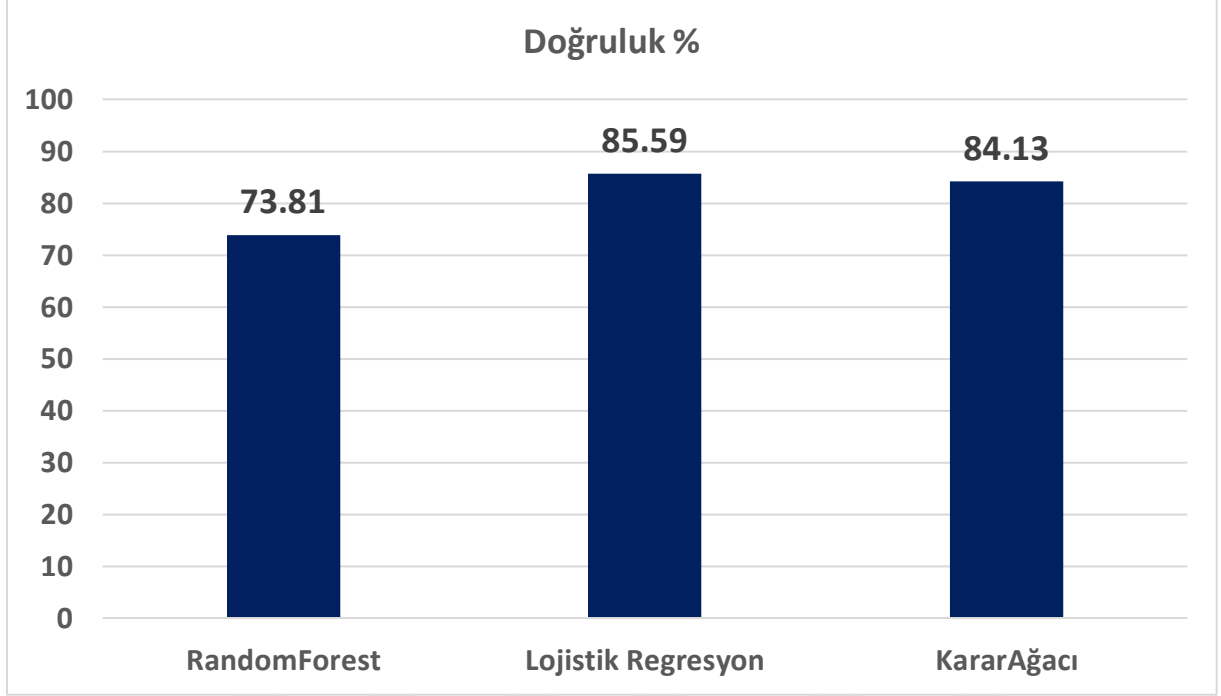
| Deęişken | P      | Odds Oranı | 95 % Güven Aralığı |
|----------|--------|------------|--------------------|
| Yaş      | < ,001 | 1,16       | (1,08 – 1,27)      |
| Cinsiyet | 0,387  | 2,26       | (3,63 – 1,55)      |
| WBC      | 0,018  | 4,84       | (1,63 – 27,6)      |
| RBC      | 0,806  | 0,54       | (0,74 – 0,43)      |
| HGB      | 0,011  | 2,46       | (1,57 – 8,18)      |
| HCT      | 0,013  | 0,33       | (< ,001– 2,28)     |
| MCV      | 0,650  | 1,66       | (0,18 – 16,4)      |
| MCH      | 0,661  | 0,23       | (< ,001- 1,38)     |
| MCHC     | < ,001 | 5,44       | (3,6 – 7,5)        |
| PLT      | 0,166  | 0,95       | (0,89 - 1,01)      |
| PDW      | 0,100  | 0,80       | (6,12 - 1,03)      |
| MPV      | 0,232  | 3,69       | (6,85 – 1,97)      |
| PCT      | 0,182  | 0,99       | (1,38 - 0,92)      |
| NEUT     | 0,063  | 0,27       | (0,04 – 0,94)      |
| NEUT.    | < ,001 | 1,53       | (1,22 – 2,10)      |
| LYM      | 0,041  | 1,33       | (1,07 - 1,69)      |
| LYM.     | < ,001 | 1,24       | (0,08 - 1,52)      |
| MONO     | 0,045  | 0,007      | (< ,0001– 0,2)     |
| MONO.    | 0,042  | 2,47       | (1,10 – 3,78)      |
| EO       | 0,059  | 0,37       | (0,06 -1,97)       |
| EO.      | 0,010  | 5,43       | (2,32 - 24,7)      |
| BASO     | < ,001 | 2,71       | (2,0 – 4,0)        |
| BASO.    | 0,297  | 4,06       | (0,32 – 8,34)      |

Yaptığımız istatistiksel analizler sonucunda Tip 2 diyabet hastalığını gösteren en önemli deęişkenler aşağıdaki Tablo 4.4’de özetlenmiştir. Tek deęişkenli analizde anlamlı olmayan şu sayısal deęişkenler tabloya alınmamıştır: WBC, RBC, HCT, MCV, PLT, PDW, MPV, PCT, NEUT., LYM, LYM., MONO., EO. Analizlerde anlamlı olan şu deęişkenler tabloya alınmıştır: Yaş, HGB, MCHC, BASO, EO, MONO.

**Tablo 4.4** Tek deęişkenli ve çok deęişkenli analizler özet tablosu

| Deęişken | Tek Deęişkenli |               |            |        | Çok Deęişkenli |        |
|----------|----------------|---------------|------------|--------|----------------|--------|
|          | Diyabetli Grup | Kontrol Grubu | Odds Oranı | p      | Odds Oranı     | p      |
| YAŞ      | 37             | 44            | 1,02       | <0,001 | 1,16           | < ,001 |
| HGB      | 14,54          | 13,73         | 0,94       | 0,019  | 2,46           | 0,011  |
| MONO     | 0,61           | 0,54          | 0,71       | 0,082  | 0,007          | 0,045  |
| MCHC     | 34,14          | 32,85         | 0,91       | <0,001 | 5,44           | < ,001 |
| BASO     | 0,04           | 0,02          | 0,99       | <0,001 | 2,71           | < ,001 |
| EO       | 0,24           | 0,12          | 0,86       | <0,001 | 0,37           | 0,059  |

Veriler ile yapılan tek ve çok deęişkenli analiz alışmasından sonra gerekleřtirilen sınıflandırma işlemlerinin sonuçları yer almaktadır. Sonuçlar deęerlendirilirken Şekil 4.1.'de görölen performans ölçütleri, 24 öznitelik içeren orijinal veri seti üzerinde lojistik regresyon, rastgele orman ve karar ağacı modelleri kullanılarak sınıflandırıldığında elde edilen deęerlerdir.

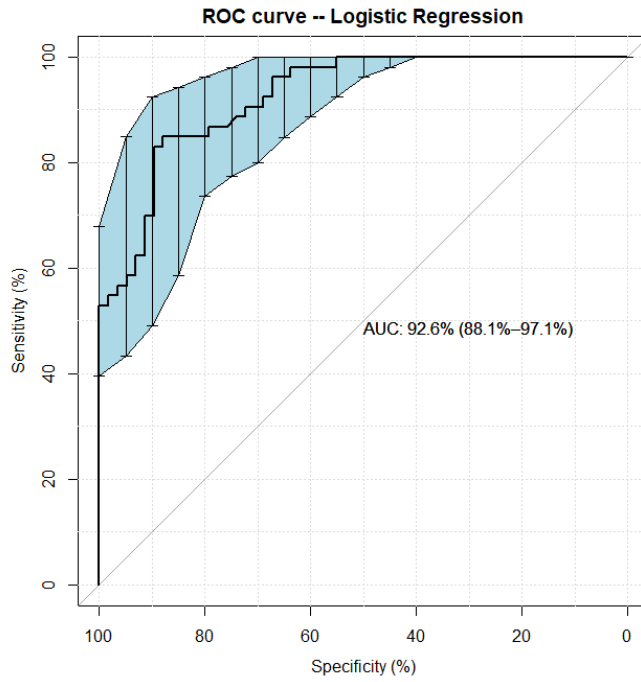


Şekil 4. 1 Makine Öğrenmeleri Accuracy Ölçümleri

Diyabetik veri seti üzerinde Lojistik Regresyon sınıflandırma teknięi uygulanmıştır. Lojistik Regresyon'un doęruluęu %85,59'dur. Ölçüm sonuçları Tablo 4.3'te gösterilmektedir. Bu modelin ROC AUC grafięi Şekil 4.2'te sunulmuştur. Lojistik Regresyon algoritması %92.6 ROC AUC başarıyla ve 88.1-97.1 güven aralığı ile sınıflandırmayı gerekleřtirmiştir. Ayrıca birden fazla bağımsız deęişkenin (girdinin) arasında doęrusal bir ilişki (bağıntı) olması durumunu incelemek için multikollinearite testleri (VIF, Variance Inflation Factor) yapılmış ve bazı deęişkenlerin VIF deęerinin yüksek olduęu görölmüştür. Bundan dolayı deęişkenlerin birer birer kullanılmasının yanıltıcı olabileceęi unutulmamalıdır.

Tablo 4. 5 Lojistik Regresyon Sınıflandırması Karmaşıklık Matrisi

| Lojistik Regresyon  | Gerçek Sınıf          |                      |                       |
|---------------------|-----------------------|----------------------|-----------------------|
| Tahmin Edilen Sınıf |                       | <i>Kontrol Grubu</i> | <i>Diyabetli Grup</i> |
|                     | <i>Kontrol Grubu</i>  | 44                   | 7                     |
|                     | <i>Diyabetli Grup</i> | 9                    | 51                    |



Şekil 4.2 Lojistik Regresyon ROC AUC

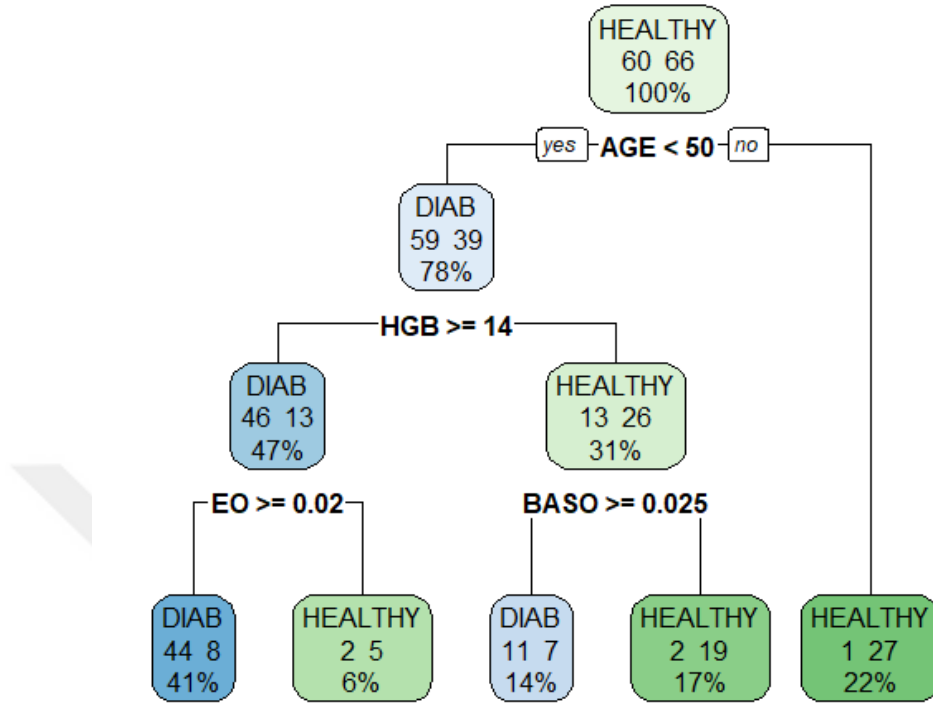
#### 4.2. Karar Ağacı

R Programlama ortamında veri setine sağlıklı ve diyabetleri ayırmak için (rpart) karar ağacı uygulanmıştır. Modelimize daha iyi bir performans kazandırmak için karmaşıklık tablosundan yararlanılmıştır. Böylelikle ağacın farklı düzeylerindeki tahmin hatalarından minimumu seçilmiştir. Şekil 4.3’de görülen karar ağacı aynı veri setinde prediktif olarak kullanıldığında %84,13 Doğruluk göstermiştir. Ağacın en alt sol yaprağında Yaşı 50 nin altında, HGB $\geq$  14 olan ve EO $\geq$ 0.02 olan 52 bireyin 44 ü Diyabetli 8 ise sağlıklıdır. En alt sağdaki yaprakta ise 50 yaş ve üzeri 28 bireyin 27si sağlıklıdır.

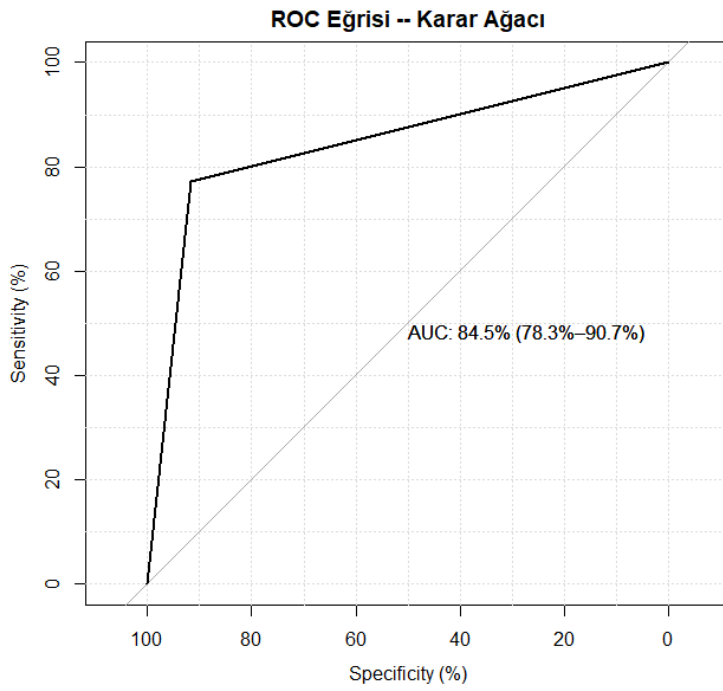
Tablo 4.6 Karar Ağacı Sınıflandırması Karmaşıklık Matrisi

| Karar Ağacı         |                | Gerçek Sınıf  |                |
|---------------------|----------------|---------------|----------------|
|                     |                | Kontrol Grubu | Diyabetli Grup |
| Tahmin Edilen Sınıf | Kontrol Grubu  | 51            | 5              |
|                     | Diyabetli Grup | 15            | 55             |

### Decision Tree



Şekil 4.3 Diyabetlilerle sağlıklıları ayırmak için oluşturulan Karar Ağacı



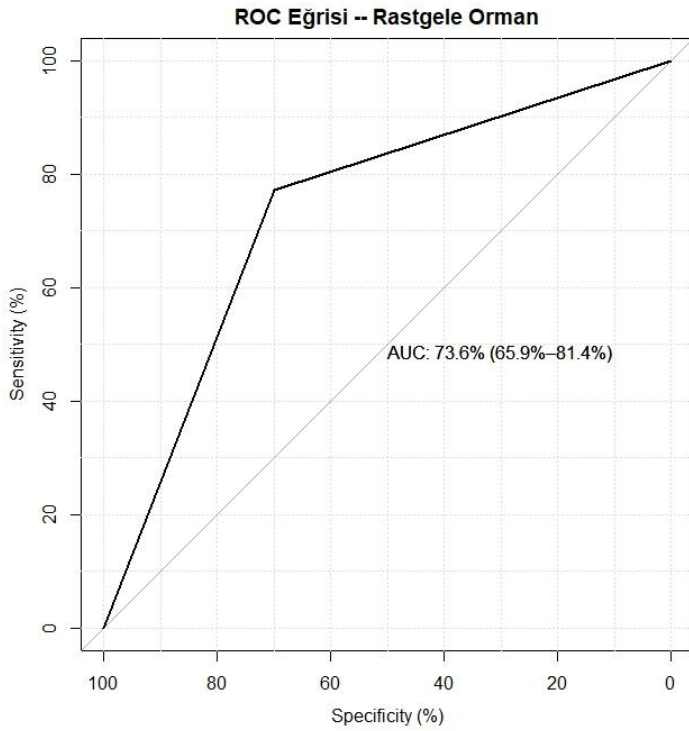
Şekil 4.4 Karar Ağacı ROC AUC

### 4.3. Rastgele Orman

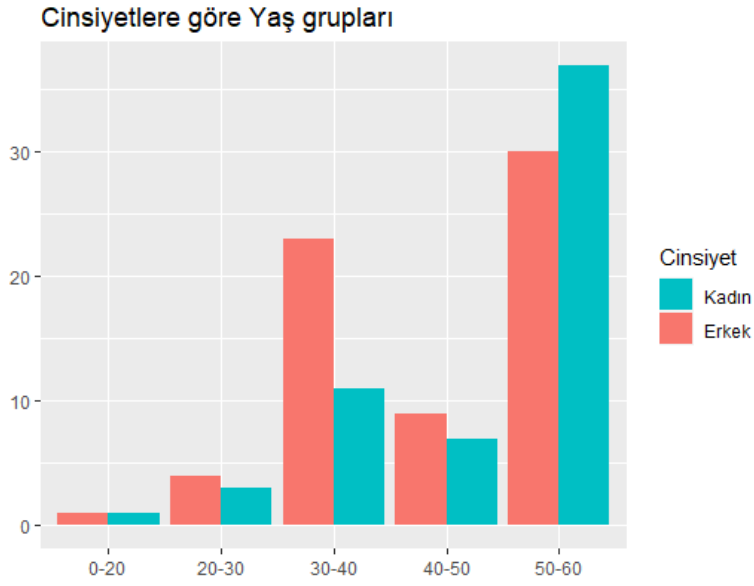
Kullanılan karar ağacı sayısı (ntree) en iyi performansı sebebiyle 30 seçilmiştir. Rastgele ormanın doğruluğu 73,81'dur. Sonuçlar Tablo 4.7'de gösterilmektedir. Bu modelin ROC AUC grafiği Şekil 4.5 te sunulmuştur.

**Tablo 4.7** Random Forest Sınıflandırması Karmaşıklık Matrisi

| Random Forest       |                       | Gerçek Sınıf         |                       |
|---------------------|-----------------------|----------------------|-----------------------|
|                     |                       | <i>Kontrol Grubu</i> | <i>Diyabetli Grup</i> |
| Tahmin Edilen Sınıf | <i>Kontrol Grubu</i>  | 42                   | 18                    |
|                     | <i>Diyabetli Grup</i> | 15                   | 51                    |



**Şekil 4.5** Rastgele Orman ROC AUC



**Şekil 4.6** Cinsiyetlere Göre Yaş Dağılım Grafiği

Cinsiyetlere göre yaş dağılım grafiği Şekil 4.6’de gösterilmiştir. Kadınlar turkuaz, erkekler ise kırmızı renk ile belirtilmiştir. Toplanan verilerimizde 0-20, 21-30, 31-40lı yaş aralığında erkeklerin daha fazla, 51-60 aralığında ise kadınların daha fazla olduğu görülmektedir.

## 5. TARTIŞMA

Bu çalışmada Türkiye'nin farklı bölgelerindeki özel hastanelerin veri tabanlarından toplanan verileri kullanarak, T2DM'nin erken tanısında makine öğrenme tekniklerinin karşılaştırılmasına odaklanmaktadır ve PIMA veri setine kullanan çalışmalara göre önemli ölçüde benzersizdir.

Çalışmamıza katılan bireylerin cinsiyete göre yaş dağılımı Şekil 4.6'de gösterilmiştir. Çalışmamızdaki diyabetli bireylerin %63'ü erkeklerden oluşmaktadır. Tigga ve Garg'ın farklı makine öğrenme algoritmaları kullanarak Tip 2 diyabet riskini tahmin ettikleri çalışmalarında, grubun %60.9'unu erkekler oluşturmuştur (Neha Prerna Tigga, 2020). Uluslararası diyabet vakfının verilerine göre, 20-79 yaş arasındaki kadınlardaki diyabet hastalığının yaygınlığı, erkeklerdeki yaygınlıktan biraz daha düşük olduğu belirtilmiştir (10.2% karşı 10.8%). 2021 yılında, diyabet hastalığı ile yaşayan erkek sayısı, kadınlardan 17.7 milyon daha fazladır (IDF, 2021). Erkekler ve kadınlardaki diyabet hastalığının yaygınlığı Şekil 2.1'de gösterilmiştir.

Çalışmamıza katılan diyabetli hastalarda WBC değeri ortalaması  $8,5 \times 10^3/\text{mm}^3$  bulunmuştur. Park JM ve arkadaşları tarafından yapılan bir çalışmada, 10 yıl boyunca gözlemlenen obez ve tip 2 diyabet hastası olmayan bireylerde, takip süresi boyunca T2DM geliştiren bireylerin  $7,4 \times 10^3/\text{mm}^3$  ve üstü WBC değerine sahip olduğu görülmüştür (Park JM L. H., 2021). Çalışmamıza dahil edilen T2DM hastalarının WBC değerleri literatürde yapılan çalışmalarla benzerlik göstermektedir. Yüksek beyaz kan hücresi sayısı ile tip 2 diyabet riski arasındaki ilişkiyi inceleyen sistematik bir inceleme, WBC değerinin tip 2 diyabet riski ile ilişkili olduğunu göstermiştir (Effrossyni gkrania-klotsas, 2010). Çalışmamızda, WBC değişkenimizde gruplar arasında anlamlı bir fark bulunamamıştır. Bunun sebeplerinden biri çalışmamızın küçük bir veri seti olması olabilir. İkinci olarak, kontrol grubundaki hastalarda bakteriyel enfeksiyon ya da kardiyovasküler hastalıklar WBC değerinde yükseliğe sebep olmuş olabilir.

Çalışmamızda T2DM hastalarına kıyasla kontrol grubunun bazofil (BASO) ve monosit (MONO) değeri daha düşüktür ve anlamlı bir fark olduğu görülmüştür. Etiyopya'da yürütülen bir çalışmada da, tip 2 diyabetik yetişkin hastaların hematolojik parametreleri kullanılmış ve bazofil değerinin T2DM hastalarında daha yüksek olduğu görülerek anlamlı bir fark olduğu bildirilmiştir (Mesay ArkewI, 2021). Başka bir çalışmada ise, tip 2 diyabetli bireylerin monosit sayısının kontrol grubuna kıyasla anlamlı derecede yüksek olduğu ve bu artışın insülin direnci ile ilişkili olduğu bulunmuştur (Katherine Esposito, Palo, & Ciotola, 2003). Son olarak, başka

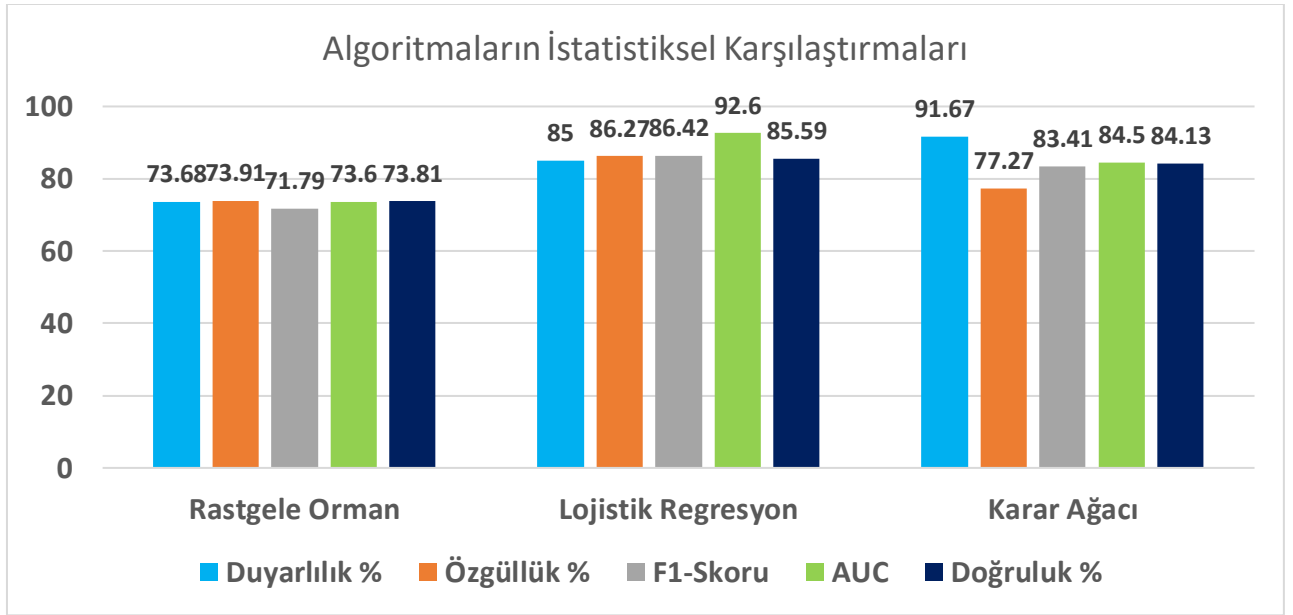
bir çalışmada, tip 2 diyabetli bireylerin monosit sayısının sağlıklı kontrol grubuna kıyasla anlamlı derecede yüksek olduğu ve bu artışın diyabetik nefropati ile ilişkili olduğu bulunmuştur (Wanghong Xu, 2014).

Çalışmamızda T2DM hastalarına kıyasla kontrol grubunun ortalama eritrosit hemoglobinin konsantrasyonu (MCHC) değerlerinin daha düşük ve anlamlı bir fark olduğu gördük. Literatürde MCHC sayısında anlamlı derecede farklılık gösteren birçok çalışma yapılmıştır. Ancak, bu çalışmaların sonuçları birbirinden farklılık gösterebilir ve araştırmacılar arasında farklı görüşler olabilir. Örneğin, bir çalışmada MCHC değerlerinin tip 2 diyabetli bireylerde kontrol grubuna kıyasla anlamlı derecede yüksek olduğu bulunmuştur. Bununla birlikte, başka bir çalışmada ise MCHC değerlerinin tip 2 diyabetli bireylerde kontrol grubuna kıyasla anlamlı derecede düşük olduğu saptanmıştır (Ebesunun, 2016 ). Diğer çalışmalarda ise, MCHC değerlerinde anlamlı bir farklılık gözlenmemiştir (Burcu DİKEÇ GÖKÇE, 2019).

Değişkenlerin Tip 2 diyabet hastalığına etkisini anlamak için tek değişkenli ve çok değişkenli istatistiksel analiz çalışmalarını yaptık. Tek değişkenli analiz çalışmamızın sonunda her bir değişkenin, Tip 2 diyabetli grup ve Kontrol grubuna göre odds oranı ve P değerleri hesaplanarak Tablo 4.2’de özetledik. Bu analiz sonucuna göre tip 2 diyabet hastalığına etki eden değişkenler şöyledir; Yaş, HGB, MCH, MCHC, BASO, EO, MONO.

Aynı veri seti ile 24 değişkenin Tip 2 diyabet hastalığını nasıl etkilediğini ortaya koymak amacıyla yaptığımız çok değişkenli (lojistik regresyon) analiz sonuçlarını Tablo 4.3’de, en önemli değişkenleri ise Tablo 4.4’te özetledik. Çalıştığımız veri setimizde Tip 2 diyabet hastalığını tahmin etmekte anlamlı değişkenlerden tam kan sayımı sonuçları arasında WBC, HGB, HCT, NEUT., LYM., MONO, EO, BASO, MCHC bulduk. Benzer şekilde Türkiye’de başka bir çalışmada lökosit, lenfosit ve nötrofil sayıları diyabetik hastalarda anlamlı bir şekilde yüksektir (Muhammed Kizilgul E. S., 2018). Mesay ve arkadaşları da tip 2 diyabetli hastalarda, eozinofil sayımlarının kontrol grubuna kıyasla anlamlı olarak daha yüksek olduğunu bulmuştur (Mesay ArkewI, 2021).

Yaptığımız çalışma sonucunda toplanan veriler üzerinde makine öğrenmesi yöntemleri test edilmiştir. Gerçekleştirdiğimiz çalışmada Karar Ağacı algoritması diğer makine öğrenmesi algoritmalarından daha fazla bir doğruluk ile sınıflandırmayı gerçekleştirmiştir.



Şekil 5.1 Algoritmaların İstatistiksel Karşılaştırması

Tablo 5.1 Makine Öğrenmeleri Performans Ölçümleri Karşılaştırması

|   | Algoritma          | Duyarlılık % | Özgüllük % | F1-Skoru | AUC   | Doğruluk % |
|---|--------------------|--------------|------------|----------|-------|------------|
| 1 | Rastgele Orman     | 73.68        | 73.91      | 71.79    | 73.6  | 73.81      |
| 2 | Lojistik Regresyon | 85.00        | 86.27      | 86.42    | 92.6  | 85.59      |
| 3 | Karar Ağacı        | 91.67        | 77.27      | 83.41    | 84.50 | 84.13      |

Yukarıdaki analizimiz, çeşitli makine öğrenimi modelleri ve performans açısından tahmin doğruluğu hakkında fikir vermektedir. Çeşitli sınıflandırmaların doğruluğunun karşılaştırılması Tablo 5.1’de gösterilmektedir. Uygulanan algoritmalar arasında en yüksek doğruluk % 85.59 ile Lojistik Regresyona aittir. Bu sonuç en iyi doğruluk oranı % 67.15 olan Saravananathan ve arkadaşlarının bulduğu sonuçlara göre daha iyi gözükmemektedir (Velmurugan, 2016). Ayrıca en yüksek duyarlılık sonucu %91.67 ile karar ağacı, %61.25 olarak en yüksek sonucu elde eden Sneha ve arkadaşlarına göre çok daha iyi bir sonuç elde edilmiştir (N. Sneha, 2019).

## 5.1. Kısıtlılıklar

Veri toplama aşamasında Sql veritabanından prosedür yardımıyla çekilen verilerde, hastanelerin bilgilerini kayıt altına aldığı tanım tablolarında, hastane seçimine bırakılan laboratuvar tetkik ve parametre isimlerinin belirli bir standarda sahip olmayışı veri toplama aşamasını yavaşlatmıştır.

Detaylı olarak incelenen verilerde ise ICD kodlarına göre Tip 2 diyabet tanısı konulmuş ancak hastaların hastane ziyaret ettiği verilerde herhangi bir tedavinin başlamadığı ve farklı ICD kodlarının girildiği görülmüştür. Bu sebeplerden ötürü kullanıcının yanlış ICD tanısı girmesi ile veri kalitesini azalttığı söylenebilmektedir.

Çalışmamızda kontrol grubuna ait veriler alınırken ICD kodlarından yararlanılarak hiçbir diyabet tanısının konulmamış ve aynı hastaneye birçok defa gelmiş hastalara ait tam kan sayımı sonuçları alınmış olmasına rağmen, bu hastaların ait diğer hastalıkları tam kan sayımı değerlerimizde farklılıklara neden olabilir.

Çalışmamızın ana kısıtlılığı bir bireyin tip 2 diyabet tanısının ilk olarak özel hastane dışındaki devlet hastanelerinde veya başka bir hastanedeki uzman hekimlerce konulmuş olması ve ardından hastanın tedavisinin devamı niteliğinde özel hastaneyi ziyaret edişi ile hastalığın ilk tanısına ait tam kan sayımı sonucu verisine erişimimiz olmaması nedeniyle daha az hasta verisine ulaşmamıza sebep olmuştur.

Diğer bir kısıtlılık Türkiye’de bakanlığa bağlı ve gönderim zorunluluğu olan Medula ve USS veri paketlerinde zorunlu alanlar dışında seçime bağlı olan alanların hastaneler tarafından doldurulması daha çok veriye ulaşılması açısından engel teşkil etmektedir. Bunlara örnek olarak e-Nabız’a gönderimi zorunlu olmayan 215 sistem kodlu Diyabet veri seti paketi içerisinde ilk tanı tarihi, boy ve kilo bilgileri gibi bilgiler sistemde zorunlu tutulmuş olmasından kaynaklı hastaneler tarafından veri girişi yapılmış olup daha çok veri erişimine sebep olabilirdi.

## 6. SONUÇ VE ÖNERİLER

Diyabet heterojen bir hastalık grubudur. Kandaki kronik glikoz yükselmesi ile karakterizedir. Amerikan diyabet derneğinin ana sloganı “Diyabeti önlemek, iyileştirmek ve diyabetten etkilenen tüm insanların yaşamlarını iyileştirmek” dir (Basevi V, 2011). Tüm dünyada insanların yaşamlarını desteklemek için sınıflandırma tekniklerini geliştirerek prediktif analizlerle diyabetin komplikasyonlarını erken aşamada tespit etmeye ve önlemeye çalışılmaktadır.

Çalışmamızın en önemli özelliği Tip 2 diyabet hastalığının henüz farkında olmayan ve bu sebeple herhangi bir sağlık hizmetine başvurmamış, başka bir nedenle tam kan sayımı testi yapılan bireylerin kan testi sonucuna göre Tip 2 diyabet hastalığının erken dönemde tanısının konulmasına yardımcı olmasıdır. Bu amaçla veri toplama aşamasında HBA1C ve açlık kan şekeri gibi değişkenlerimize ait verilerimiz olsa da yapılan analizlerimizde bu değişkenlere yer vermedik. Yaptığımız analizler doğrultusunda HGB, MCHC, BASO ve EO değerlerinde önemli sonuçlar elde edilmiştir. Kullanılan verilerin hazır bir veri setinden alınmamış, Türkiye’deki farklı hastanelerden toplanmış olması çalışmamıza özgünlük katmıştır.

Karar Ağacı algoritması ve en yüksek değeri veren Lojistik Regresyon sırasıyla %77,27 ve %86,27’lik özgüllük ile diyabetik verilerin analizi için en iyisidir. Lojistik Regresyon, Rastgele orman ve Karar ağacı teknikleri mevcut yöntemden sırasıyla %85,59 , %73,81 ve %84,13 doğruluk vermektedir. Karar ağacı algoritması bireylerin tip 2 diyabet hastalığına sahip olmasını en iyi tespit eden algoritma olmuştur. Bu değerin tespit edilmesinde kullanılan değerlendirme ölçütü ‘Duyarlılık’ değeridir. Lojistik Regresyon algoritması bu değeri % 85,00 olarak bulmuşken, Rastgele Orman % 73,68 olarak bulmuş ve Karar ağacı ise % 91,67 değeri ile en iyi sonucu elde etmiştir.

Bu çalışmada yeni bir yöntem önerilmemiş olsa da, gerçek bir veri seti üzerindeki makine öğrenmesi yöntemlerinin performansı ölçülmüştür. Gerçekleştirilen çalışma, bir deneyimin aktarılmasıdır.

- Çalışmamızın en önemli kısıtlılığı daha büyük bir veri setinin olmamasıdır. Bunun nedeni veri toplanma aşamasında istenilen kriterlere uygun hasta verisi bulmanın zorluğu idi. Ayrıca standart veri kümesindeki eksik veri örnekleri de çıkarıldı; bu durum da veri setinin küçülmesine neden oldu. Standart veri kümesindeki eksik değerleri çıkarıldı; bu, veri kümesinin boyutu küçüldükçe sonuçları etkileyebilir.

- Kan testleri parametreleri ile yapılan az sayıda çalışma mevcuttur. Çalışmamız gelecekteki çalışmalara bir temel oluşturabilir, daha büyük veri setleri ile farklı makine öğrenmesi algoritmaları uygulanarak sonuçlar karşılaştırılabilir.



## 7. KAYNAKLAR

(2022, 09 6). Pandas: <https://pandas.pydata.org/> adresinden alındı

A review. (2015). *Decision support systems for predicting diabetes mellitus*.

Abdelkader, W. N. (2021). Machine Learning Approaches to Retrieve High-Quality, Clinically Relevant Evidence From the Biomedical Literature: Systematic Review. *MIR Medical Informatics*.

Abdul Azis Abdillah, S. S. (2016). Diagnosis of Diabetes using Support Vector Machines with Radial Basis Function Kernels. *International Journal of Technology*, 849-858.

Adane T, A. F. (2021). White blood cells and platelet profiles of diabetic patients at University of Gondar specialized referral hospital: A comparative cross-sectional study. *J Clin Lab Anal.*, 35(6).

Ahmed Kareem Oleiwi, L. S. (2020). A Comparative Analysis and Risk Prediction of Diabetes at Early Stage using Machine Learning Approach. *International Journal of Future Generation Communication and Networking*, 4151–4163.

Ahmet Anık, G. Ç. (2014). Maturity-onset diabetes of the young (MODY): an update.

Ahuja, R. S. (2019). A Diabetic Disease Prediction Model Based on Classification Algorithms. *Annals of Emerging Technologies in Computing*, 44–52.

Akinsola, J. E. (2017). Supervised Machine Learning Algorithms: Classification and Comparison. *International Journal of Computer Trends and Technology*, 128-138.

Ali, N. N. (2019). Evaluation of k-nearest neighbour classifier performance for heterogeneous data sets. *SN Appl. Sci.*

Allaoui G, R. C. (2022). Longitudinal changes in blood biomarkers and their ability to predict type 2 diabetes mellitus-The Tromsø study. *Endocrinol Diabetes Metab.*, 5(2).

Allaoui, G. R. (2022). Longitudinal changes in blood biomarkers and their ability to predict type 2 diabetes mellitus—The Tromsø study. *Endocrinology, Diabetes & Metabolism*.

Alshehri, Z. (2017). The relationship between some biochemical and hematological changes in type 2 diabetes mellitus. *Biomedical Research and Therapy*, 4(11):1760.

Arooj Hussain, S. N. (2021). Prediction of Diabetes Mellitus: Comparative Study of Various Machine Learning Models. *International Conference on Innovative Computing and Communications*, 103-115.

Basevi V, D. M. (2011). American Diabetes Association. Standards of medical care in diabetes. *Diabetes Care*, S11-S61.

Bastaki, S. (2005). Diabetes mellitus and its treatment.

- Bhanu, A. B. (2015). A visual representation of machine learning-based predictive modeling.
- Bliss, M. (1984). *The Discovery of Insulin*.
- Breiman, L. (1999). Random Forests--Random Features.
- Burcu DİKEÇ GÖKÇE, A. G. (2019). Tip 2 Diyabetli Hastalarda Dapagliflozin Tedavisinin Trombosit Fonksiyonları ve İnflamasyon Üzerine Etkisi. *Uludağ Üniversitesi Tıp Fakültesi Dergisi*.
- Chahal, D. H. (2013). ID3 modification and implementation in data mining. *International Journal of Computer Applications*. , 80.
- Committee, C. D., & Ronald Goldenberg, Z. P. (2013). Definition, classification and diagnosis of diabetes, prediabetes and metabolic syndrome.
- Cover T., H. P. (1967). Nearest Neighbor Pattern Classification. *IEEE Transactions on Information Theory*, 21-27.
- Daniele Soria, J. M. (2011). A ‘non-parametric’ version of the naive Bayes classifier. *Knowledge-Based Systems*, 775-784.
- Deepthi B, S. K. (2017). A Modern Review of Diabetes Mellitus: An Annihilatory Metabolic Disorder. doi:10.21767/2469-6692.100014
- Devi, S. (2020). Logistic Regression for Employability Prediction.
- Dilip Kumar Choubey, M. K. (2020). Comparative Analysis of Classification Methods with PCA and LDA for Diabetes.
- Doddipalli, L. &. (2011). Performance Evaluation of Decision Tree Classifiers on Medical Datasets. *International Journal of Computer Applications*, 26. doi:10.5120/3095-4247
- Doupe P, F. J. (2019). Machine Learning for Health Services Researchers. *Value Health*, 808-815.
- E Laugesen, J. A. (2015). Latent autoimmune diabetes of the adult: current knowledge and uncertainty.
- Ebesunun, M. O. (2016 ). Male Sexual Dysfunction, Leptin, Pituitary and Gonadal Hormones in Nigerian Males with Metabolic Syndrome and Type 2 Diabetes Mellitus.
- Eduardo P. Costa, A. C. (2005). A Review of Performance Evaluation Measures for Hierarchical Classifiers.
- Effrossyni gkrania-klotsas, Z. Y. (2010). Differential White Blood Cell Count and Type 2 Diabetes: Systematic Review and Meta-Analysis of Cross-Sectional and Prospective Studies. *Public Library of Science*.

- Elosta A, G. T. (2012 ). Natural products as anti-glycation agents: possible therapeutic potential for diabetic complications. 8(2):92-108.
- Essam H. Houssein, M. M. (2021). Deep and machine learning techniques for medical imaging-based breast cancer: A comprehensive review. *Expert Systems with Applications*, 167.
- FARHUD, D. D. (2015). Impact of Lifestyle on Health. *Iran J Public Health*, 1442-1444.
- G. M. Foody, M. K. (2010). An evaluation of some factors affecting the accuracy of classification by an artificial neural network. *International Journal of Remote Sensing*, 799-810.
- Gkrania-Klotsas E, Y. Z.-K. (2010 ). Differential white blood cell count and type 2 diabetes: systematic review and meta-analysis of cross-sectional and prospective studies. *PLoS One*, 18;5(10).
- Han Wu, S. Y. (2018). Type 2 diabetes mellitus prediction model based on data mining. *Informatics in Medicine Unlocked*, 100-107.
- Harikumar K., K. B. (2016). A review on diabetes mellitus.
- Hasdyna, N. (2020). Improving The Performance of K-NearestNeighbor Algorithm by Reducing The Attributes ofDataset Using Gain Ratio. *Journal of Physics: Conference Series*.
- Hernandez LM, B. D. (2006). The Impact of Social and Cultural Environment on Health.
- Hetal Bhavsar, A. G. (2012). A Comparative Study of Training Algorithms for Supervised Machine Learning. *International Journal of Soft Computing and Engineering*.
- Hong Han, X. G. (2016). Variable selection using Mean Decrease Accuracy and Mean Decrease Gini based on Random Forest. doi:10.1109/icsess.2016.7883053
- Hongzhou Liu, J. L. (2022). Triglyceride to High-Density Lipoprotein Cholesterol (TG/HDL-C) Ratio, a Simple but Effective Indicator in Predicting Type 2 Diabetes Mellitus in Older Adults. *Clinical Diabetes*.
- Hssina, B. &. (2014). A comparative study of decision tree ID3 and C4.5. *International Journal of Advanced Computer Science and Applications*. doi:10.14569/SpecialIssue.2014.040203
- <https://www.rstudio.com/products/rstudio/>. (2022, 10 27). RStudio: <https://www.rstudio.com/products/rstudio/> adresinden alındı
- IDF, I. D. (2021). *IDF Diabetes Atlas 10th edition*.
- intuitive. (2023, 01 10). intuitive: <https://www.intuitive.com/en-us/about-us/company> adresinden alındı

- Islam, M. M. (2020). Likelihood Prediction of Diabetes at Early Stage Using Data Mining Techniques. *Computer Vision and Machine Intelligence in Medical Image Analysis* , 113-125.
- Islam, R. A. (2019). Binary classification techniques in machine learning: A comparative study.
- J. Maccracken, D. H. (1997). From ants to analogues. *Postgraduate Medicine*, 138-150.
- J. Pradeep Kandhasamy, S. B. (2015). Performance Analysis of Classifier Models to Predict Diabetes Mellitus. *Procedia Computer Science*, 45-51.
- Jair Cervantes, F. G.-L.-M. (2020). A comprehensive survey on support vector machine classification: Applications, challenges and trends. *Neurocomputing*.
- Jatinder Kaur, J. S. (2015). Optimizing the Accuracy of CART Algorithm by Using Genetic. *International Journal of Computer Science Trends and Technology*.
- Jazayeri A, L. O. (2020). Imputation of Missing Data in Electronic Health Records Based on Patients' Similarities. *J Healthc Inform Res.*, 295-307.
- Jian-xiong Dong, A. K. (2005). Fast SVM training algorithm with decomposition on very large data sets. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 603-618.
- Jiawei Han, M. K. (2012). *Data Mining Concepts and Techniques Third Edition*.
- Kamber, J. H. (2011). *Data Mining: Concepts and Techniques*.
- Katherine Esposito, F. N., Palo, C. D., & Ciotola, M. (2003). Cytokine Milieu Tends Toward Inflammation in Type 2 Diabetes. *Diabetes Care*.
- Kenéz, J. (1978). Who invented insulin? (Charles H. Best, Frederick Grant Banting). 119(44):2693-9.
- Khan, M. H. (2019). Epidemiology of Type 2 Diabetes – Global Burden of Disease and Forecasted Trends. *Journal of Epidemiology and Global Health* 10, 107-111.
- Kibria HB, N. M. (2022). An Ensemble Approach for the Prediction of Diabetes Mellitus Using a Soft Voting Classifier with an Explainable AI. *Sensors*, 22(19).
- Kushan De Silva, W. K. (2020). Use and performance of machine learning models for type 2 diabetes prediction in community settings: A systematic review and meta-analysis. *International Journal of Medical Informatics*, 1386-5056.
- Lai H, H. H. (2019 ). Predictive models for diabetes mellitus using machine learning techniques. *BMC Endocr Disord*.
- Lai H, H. H. (2019). Predictive models for diabetes mellitus using machine learning techniques. *BMC Endocr Disord*, 19(1):101. doi:10.1186/s12902-019-0436-6

- Lakhtakia, R. (2013). The History of Diabetes Mellitus.
- Langley, G. H. (1995). Estimating continuous distributions in Bayesian classifiers. *In Proceedings of the Eleventh conference on Uncertainty in artificial intelligence*, 338–345.
- Larabi Marie-Sainte, S. &. (2019). Current Techniques for Diabetes Prediction: Review and Case Study. *Applied Sciences*.
- Lin, C.-W. H.-J. (2002). A comparison of methods for multiclass support vector machines. *IEEE Transactions on Neural Networks*, 415-425.
- M. Razzak, S. Z. (2018). A survey on supervised learning algorithms.
- Mackerras, D. E. (2003). Does a combined program of dietary modification and physical activity or the use of metformin reduce the conversion from impaired glucose tolerance to type 2 diabetes? *Med J Aust.*, 178(4):180-1.
- Majumdar, S. K. (2001). GLIMPSES OF THE HISTORY OF INSULIN. 57-70.
- Maniruzzaman M, S. H.-B. (2018). Risk factors of neonatal mortality and child mortality in Bangladesh. *J Glob Health*.
- Maniruzzaman, M. &. (2020). Classification and prediction of diabetes disease using machine learning paradigm. *Health Information Science and Systems*, 1-14.
- Meigs, J. (2009). Multiple Biomarker Prediction of Type 2 Diabetes. *Diabetes Care*, 1346–1348.
- Mesay Arkew, T. Y. (2021). Hematological parameters of type 2 diabetic adult patients at Debre Berhan Referral Hospital, Northeast Ethiopia: A comparative cross-sectional study. *PLoS One*.
- Mesay ArkewI, T. Y. (2021). Hematological parameters of type 2 diabetic adult patients at Debre Berhan Referral Hospital, Northeast Ethiopia: A comparative cross-sectional study. *PLOS ONE*.
- Micheal O. Olusanya, R. E. (2022). Accuracy of Machine Learning Classification Models for the Prediction of Type 2 Diabetes Mellitus: A Systematic Survey and Meta-Analysis Approach. *International journal of environmental research and public health*, 19.
- MINGERS, J. (1989). An Empirical Comparison of Pruning Methods for Decision Tree Induction. 227-243.
- Mittal, M. R. (2018). A Survey on Semi-Supervised Learning.
- Mittal, S. J. (2018). Unsupervised learning algorithms: a review.
- Mohapatra, S. &. (2019). Detection of Diabetes Using Multilayer Perceptron: Proceedings of ICICA 2018.

- Muhammed Kizilgul, E. S. (2018). Components of the Complete Blood Count in Type 2 Diabetes. *Dicle Medical Journal* , 113-120.
- Muhammed Kizilgul, E. S. (2018). Yetersiz Glisemik Kontrolü Olan Tip 2 Diyabet Hastalarında Tam Kan Sayımı Parametreleri. *Dicle Medical Journal*, 113-120.
- Murugesan Thinakaran, T. C. (2021). COMPARATIVE HEMATOLOGICAL STUDY OF TYPE 2 DIABETIC AND NON-DIABETIC PATIENTS IN ERODE DISTRICT. *Indian Journal of applied research*.
- N. Sneha, T. G. (2019). Analysis of diabetes mellitus for early . *Journal of Big Data*.
- NainaMahajan, B. (2016). Analysis of Factors of Road Traffic Accidents using Enhanced Decision Tree Algorithm. *International Journal of Computer Applications*, 135 .
- Natascia Di Iorgi, F. N. (2012). Diabetes insipidus--diagnosis and management.
- Needham, J. (1959). J. Needham içinde, *Science and Civilisation in China: Volume 3, Mathematics and the Sciences of the Heavens and the Earth*. Cambridge University Press.
- Neha Prerna Tigga, S. G. (2020). Prediction of Type 2 Diabetes using Machine Learning Classification Methods. *Procedia Computer Science*, 706-716.
- Ngugi M.P, N. J. (2012). Diagnosis of Diabetes Mellitus. *International Journal of Experimental Diabetes Research*, 24-27.
- Nguyen, G. D. (2019). Machine Learning and Deep Learning frameworks and libraries for large-scale data mining: a survey. *Artif Intell Rev*, 77–124.
- Nongyao Nai-arun, R. M. (2019). Comparison of Classifiers for the Risk of Diabetes Prediction. *Procedia Computer Science*, 132-142.
- Nurrohman, A. A. (2020). Parkinson's disease subtype classification: Application of decision tree, logistic regression and logit leaf model. *AIP Conference Proceedings*. doi:10.1063/5.0007880
- Obuchowski, N. A. (2000). ROC Analysis.
- Oguzhan, G. D. (2020). Sosyoekonomik Durumun Diyabet ile İlişkisi ve Diyabete Etkisinin İncelenmesi. *urkish Journal of Diabetes and Obesity* , 71–78.
- Özge Nur Ergün, H. O. (2021). Early Stage Diabetes Prediction Using Machine Learning Methods. *European Journal of Science and Technology*, 52-57.
- Ozougwu, J. C. (2013). The pathogenesis and pathophysiology of type 1 and type 2 diabetes mellitus.
- Pal, M. (2005). Random forest classifier for remote sensing classification. *International Journal of Remote Sensing*, 217–222. doi:10.1080/01431160412331269698

- Pal, M. (2005). Random forest classifier for remote sensing classification. *International Journal of Remote Sensing*, 217-222. doi:10.1080/01431160412331269698
- Parikh R, M. A. (2008 ). Understanding and using sensitivity, specificity and predictive values. *Indian J Ophthalmol*, 45-50.
- Park JM, L. H. (2021). White Blood Cell Count as a Predictor of Incident Type 2 Diabetes Mellitus Among Non-Obese Adults: A Longitudinal 10-Year Analysis of the Korean Genome and Epidemiology Study. *J Inflamm Res*, 1235-1242.
- Park JM, L. H. (2021). White Blood Cell Count as a Predictor of Incident Type 2 Diabetes Mellitus Among Non-Obese Adults: A Longitudinal 10-Year Analysis of the Korean Genome and Epidemiology Study. *J Inflamm Res.*, 1;14:1235-12.
- Payam Refaeilzadeh, L. T. (2016). Cross-Validation. *Encyclopedia of Database Systems*, 1-7.
- Pedregosa, F. &. (2012). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*.
- Qiang Wu, D.-X. Z. (2006). Analysis of support vector machine classification. *Journal of Computational Analysis and Applications*.
- Qing-yun Dai, C.-p. Z. (2016). Research of Decision Tree Classification Algorithm in Data Mining.
- Raju, V. K. (2003). *Susruta of Ancient India*. 51:119-22.
- Ramachandran, A. (2014). Know the signs and symptoms of diabetes. 140(5): 579–581.
- Rani, K. U. (2011). Performance Evaluation of Decision Tree Classifiers on Medical Datasets. *International Journal of Computer Applications*.
- Rattray, Z. G. (2007). Semi-Supervised Learning: A Survey.
- Reduction in the Incidence of Type 2 Diabetes with Lifestyle Intervention or Metformin. (2002). *The New England Journal of Medicine*, 393–403. doi:10.1056/nejmoa012512
- Robin Prakash Mathur, A. R. (2013). Survey on Decision Tree classification algorithms for the Evaluation of Student Performance.
- S. S. Al-Sarawi, D. L. (2012). A Comparative Study of Information Gain and Gini Index for Feature Selection.
- Saloni Kumari, D. K. (2021). An ensemble approach for classification and prediction of diabetes mellitus using soft voting classifier.
- Sarker, I. (2021). Machine Learning: Algorithms, Real-World Applications and Research Directions.

- Schmidt MI, D. B. (1999). Markers of inflammation and prediction of diabetes mellitus in adults (Atherosclerosis Risk in Communities study): a cohort study. *Lancet*.
- Sedat METLEK, K. K. (2020). *MAKİNE ÖĞRENMESİNDE, TEORİDEN ÖRNEK MATLAB UYGULAMALARINA KADAR DESTEK VEKTÖR MAKİNELERİ*.
- Shahbazian, H. (2013). World diabetes day.
- Shahid Mohammad Ganie, M. B. (2020). Comparative analysis of various supervised machine learning algorithms for the early prediction of type-II diabetes mellitus. *International Journal of Medical Engineering and Informatics* .
- Siang Yong Tan, J. M. (2017). Frederick Banting (1891–1941): Discoverer of insulin. 2.
- Sim, J. W. (2005). The Kappa Statistic in Reliability Studies: Use, Interpretation, and Sample Size Requirements. *Physical Therapy* , 257–268.
- Smriti Gupta, H. K. (2021). Classification of Diabetes Using Naïve Bayes and Support Vector Machine as a Technique. *Springer Singapore*.
- Sneha, N. G. (2019). Analysis of diabetes mellitus for early prediction using optimal features selection. *Journal of Big Data*.
- Souad Larabi-Marie-Sainte, L. A. (2019). Current Techniques for Diabetes Prediction: Review. *applied science*.
- Srinivasan, K. (2005). Plant foods in the management of diabetes mellitus: Spices as beneficial antidiabetic food adjuncts.
- Sumbal Malik, S. H.-S. (2020). Comparative Analysis of Machine Learning Algorithms for Early Prediction of Diabetes Mellitus in Women. *Modelling and Implementation of Complex Systems, Proceedings of the 6th International Symposium*, 95-106.
- Topol, E. (2019). High-performance medicine: the convergence of human and artificial intelligence. *Nature Medicine*, 25, 44–56. doi:10.1038/s41591-018-0300-7
- Topol, E. J. (2019). High-performance medicine: the convergence of human and artificial intelligence. 44–56.
- Torino, A. O. (1989). Giovanni Martinotti and experimental pancreatectomy, a century ago.
- Türkiye Diyabet Vakfı*. (2022). Türkiye Diyabet Vakfı: <https://www.turkdiab.org/> adresinden alındı
- (2020). *Türkiye’de Diyabette Çözüm Yolları*. Türkiye diyabet vakfı.
- Velmurugan, K. S. (2016). Analyzing Diabetic Data using Classification. *Indian Journal of Science and Technology*, 1-6.
- Verma, M. D. (2017). Decision tree based machine learning algorithms: A review.

- Wang, T. J. (2011). Assessing the Role of Circulating, Genetic, and Imaging Biomarkers in Cardiovascular Risk Prediction. *Risk Prediction in Cardiovascular Medicine*, 551–565.
- Wanghong Xu, Y. Y. (2014). Dietary fiber intake and risk of type 2 diabetes: a dose-response analysis of prospective studies.
- Wen Zhu, N. Z. (2010). Sensitivity, Specificity, Accuracy, Associated Confidence Interval and ROC. *Health Care and Life Sciences*.
- WHO. (2010). WHO: <https://www.who.int/europe/news-room/fact-sheets/item/a-healthy-lifestyle---who-recommendations> adresinden alındı
- Wiki. (2022, 09). Wiki : [https://en.wikipedia.org/wiki/C4.5\\_algorithm](https://en.wikipedia.org/wiki/C4.5_algorithm) adresinden alındı
- wikipedia. (2023, 01 10). wikipedia: <https://tr.wikipedia.org/wiki/Cezer%C3%AE> adresinden alındı
- wikipedia. (2023, 01 0). wikipedia: [https://en.wikipedia.org/wiki/Leonardo%27s\\_robot#cite\\_note-nih-1](https://en.wikipedia.org/wiki/Leonardo%27s_robot#cite_note-nih-1) adresinden alındı
- wikipedia. (2023, 01 10). wikipedia: [https://tr.wikipedia.org/wiki/John\\_McCarthy\\_\(bilgisayar\\_bilimci\)](https://tr.wikipedia.org/wiki/John_McCarthy_(bilgisayar_bilimci)) adresinden alındı
- Wilson PW, M. J. (2007). Prediction of incident diabetes mellitus in middle-aged adults: the Framingham Offspring Study. *Arch Intern Med.*, 1068-1074.
- Wittemans, L. L. (2018). Prioritising Risk Factors for Type 2 Diabetes: Causal Inference through Genetic Approaches. *Current Diabetes Reports*. doi:10.1007/s11892-018-1009-1
- Wondifraw, H. (2015). Classification, Pathophysiology, Diagnosis and Management of Diabetes Mellitus.
- World Health Organization. (2022, 09 03). WHO: [https://www.who.int/health-topics/diabetes#tab=tab\\_1](https://www.who.int/health-topics/diabetes#tab=tab_1) adresinden alındı
- Wu, Y. F. (2018). Development and Application of Artificial Neural Network. *Wireless Pers Commun* , 1645–1656 .
- Xu, S. (2016). Bayesian Naive Bayes classifier to text classification. *Journal of Information Science*, 48–59.
- Yuvaraj, N. S. (2019). Diabetes prediction in healthcare systems using machine learning algorithms on Hadoop cluster. *Cluster Computing* , 22, 1–9.
- Zaric BL, O. M.-M. (2019). Drug Delivery System in the Treatment of Diabetes Mellitus. 25(2):166-173.

Zhilbert Tafa, N. P. (2015). An intelligent system for diabetes prediction. *Mediterranean Conference on Embedded Computing*.

