

Makine Öğrenmesi ile Erken Evre Diyabet Riskinin Sınıflandırılması

Kağan Güner

kagan.guner@gazi.edu.tr

Bilgisayar Mühendisliği, Teknoloji Fakültesi, Gazi Üniversitesi, Ankara, Türkiye

Davronbek Abdurazzokov

23181616403@gazi.edu.tr

Bilgisayar Mühendisliği, Teknoloji Fakültesi, Gazi Üniversitesi, Ankara, Türkiye

Anahtar Kelimeler

Erken Evre Diyabet,
Makine Öğrenmesi,
Sınıflandırma Algoritmaları

Özet: Diyabet, yaygın ve potansiyel olarak ölümcül bir hastalıktır. Dünya genelinde milyonlarca insan bu hastalıkla mücadele etmekte olup diyabet, hastaların yaşam kalitesini ciddi şekilde etkilemektedir. Erken müdahale ve tedavi ile diyabetin olumsuz etkileri büyük ölçüde azaltılabilir, hastaların yaşam standartları iyileştirilebilir. Ancak, çoğu durumda diyabetin tanısı yıllar süren bir süreç gerektirmektedir. Diyabetin erken teşhisi için, mevcut hastaların verilerinden yararlanarak makine öğrenmesi algoritmaları uygulanabilir. Bu yaklaşımla, kan testi veya glukoz ölçümü gibi tıbbi prosedürlere gerek kalmadan, diyabet riski taşıyan bireyler belirlenebilir. Bu makalede hedeflenen, diyabet tanısında kullanılabilecek bir makine öğrenmesi modelinin geliştirilmesidir. Araştırmada, 520 hastaya ait 16 farklı veri özelliği kullanılarak yedi farklı makine öğrenmesi algoritması uygulanmış, modellerin performansı doğruluk, kesinlik, duyarlılık ve F skoru gibi metrikleriyle değerlendirilmiştir. Ayrıca, veri setindeki özelliklerin diyabet teşhisindeki öncelikli etkileri araştırılmıştır. Geliştirilen modellerin tamamı belirli düzeylerde başarı elde etmiştir. En düşük doğruluk oranı, Lineer Regresyon algoritması kullanılarak %60 sınıflandırma başarıyla elde edilmiştir. En yüksek başarı ise Rastgele Orman algoritması ile elde edilmiştir; bu modelin doğruluğu %99, kesinliği %99, hassasiyeti %99 ve F skoru %99 olarak ölçülmüştür. Elde edilen sonuçlar, geliştirilen sınıflandırma modelinin diyabet teşhisinde bir test aracı olarak kullanılabileceğini göstermektedir. [1] [2] [3] [4]

Keywords

Early Stage Diabetes,
Machine Learning,
Classification Algorithms

Abstract: Diabetes is a widespread and potentially fatal disease. Millions of people around the world struggle with it, and the condition seriously diminishes patients' quality of life. With early intervention and treatment, the adverse effects of diabetes can be greatly reduced and patients' living standards improved. Yet in many cases, diagnosing diabetes is a process that can take years. To enable earlier detection, machine-learning algorithms can be applied to data from existing patients. In this way, people at risk of diabetes can be identified without resorting to medical procedures such as blood tests or glucose measurements. This article aims to develop a machine-learning model suitable for diagnosing diabetes. Using 16 distinct features from 520 patients, the study applied seven different machine-learning algorithms and evaluated their performance in terms of accuracy, precision, recall (sensitivity), and F-score. The influence of each feature on diabetes diagnosis was also examined. All of the models achieved a certain level of success. The lowest performance came from the Linear Regression algorithm, which reached 60 % accuracy. The highest performance was obtained with the Random Forest algorithm, which scored 99 % on accuracy, precision, recall, and F-score alike. These results indicate that the developed classification model could serve as a valuable testing tool for diabetes diagnosis. [1] [2] [3] [4]

1. Giriş

Diyabet, pankreasın yeterince insülin üretmemesi ya da vücudun üretilen insülini etkili şekilde kullanamaması durumunda ortaya çıkan kronik bir hastalıktır. Yoğun susuzluk, aşırı yorgunluk hissi, normalden fazla idrara çıkma ihtiyacı belirtileri irdelenmez ve önlem alınmaz ise önce damarlarda, hastalık ilerlediğinde ise şekerin vücuda zarar veren toksik etkisi böbrek yetmezliği, kalp hastalıkları, görme kaybı gibi ciddi komplikasyonlara yol açabilir. Erken teşhisle birlikte tanısı konulan bireylerde diyet, egzersiz ve gerektiğinde ilaç tedavisiyle kan şekeri kontrol altına alınarak daha büyük hasarların önüne geçilmesini, uzun vadede bireyin yaşam kalitesini artırmada önemli rol oynar. [1] [2]

Uluslararası Diyabet Federasyonu'nun (IDF) 2021 verilerine göre, küresel çapta diyabetin yayılışı hızla artmaktadır. Dünyada diyabetle yaşayan 20-79 yaş arası yetişkin sayısı 537 milyon iken, bu sayının 2045 yılında 783 milyona ulaşması beklenmektedir. Bu artışın en büyük nedenleri arasında obezite, hareketsizlik, sağlıklı beslenme ve yaşlanma bulunmaktadır. Sağlık Bakanlığı'nın 2018 yılı verilerine göre, Türkiye, OECD ülkeleri arasında diyabet sıklığında ikinci sıradadır. Türkiye Diyabet Vakfı'nın raporuna göre, ülkede yaklaşık 7 milyon diyabet hastası bulunmakta ve diyabetin görülme sıklığı son 10 yılda iki katına çıkmıştır. Ayrıca Türkiye, Avrupa ülkeleri arasında diyabetin en hızlı arttığı ülke olarak kaydedilmiştir. Bu veriler, erken teşhis ve temeli bilime dayanan önleme yöntemlerinin önemini ortaya koymaktadır. [5] [6] [7]

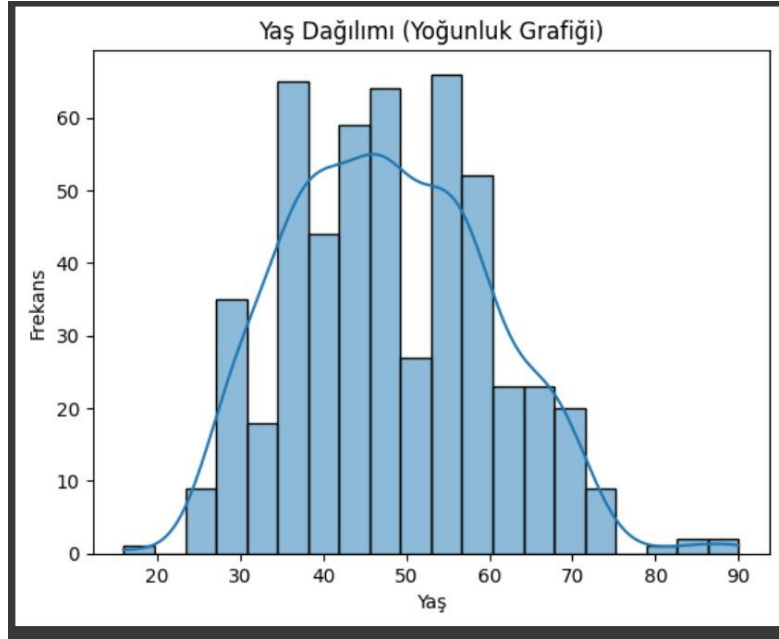
Bu makale, semptom temelli diyabet taramasını çoklu ML algoritmalarıyla sistematik olarak test eder. Bu bağlamda rapor şu araştırma sorularını ele alır: (i) Semptomlara dayalı verilerle diyabetin erken teşhisi ne kadar başarılı bir şekilde yapılabilir? (ii) Kullanılan makine öğrenme yöntemlerinden hangisi veri setimiz için en yüksek doğruluk sağlar? (iii) Veri yapısına bağlı olarak modeller arasında ne tür performans farkları gözlemlenir? Bu soruları yanıtlamak için, literatürde yaygın olarak kullanılan ancak sınırlılıklar içeren veri setleri (örneğin yalnızca kadınlardan oluşan Pima Indians veri seti) yerine, farklı yaş ve cinsiyet gruplarından bireyleri kapsayan ve semptom bazlı özellikler içeren **Early Stage Diabetes Risk Prediction Dataset** tercih edilmiştir. Araştırmada; K-en yakın komşu (K-Nearest Neighbors), Karar Ağacı (Decision Tree), Naive Bayes, Doğrusal Regresyon (Linear Regression), Lojistik Regresyon (Logistic Regression) ve Rastgele Orman (Random Forest) algoritmaları uygulanmış ve performansları doğruluk, kesinlik, duyarlılık ve F1 skoru gibi ölçütler üzerinden değerlendirilmiştir.

[7] [8] [9]

2. Materyal ve Yöntem

2.1. Veri Setinin Düzenlenmesi

Veri seti, Hindistan merkezli sağlık kuruluşlarından toplanan 520 bireyin sağlık semptomlarını içermektedir. Verisi toplanan bireylerin en düşük yaşı 16, en yüksek yaşı 90; 328 erkek, 192 kadın olduğu tespit edilmiştir. Veri kümesinde; yaş, cinsiyet, açlık kan şekeri düzeyi, kaşıntı, polidipsi, yorgunluk, bulanık görme, obezite gibi 17 adet semptom niteliği içermektedir. Hedef değişken ise bireyin diyabet riskine sahip olup olmadığını (class: Positive / Negative) belirtmektedir.



Şekil i. Çalışmaya Katılanların Yaş Dağılımı

Veri ön işleme aşamasında, öncelikle “Yes/No”, “Male/Female” gibi kategorik değerler binary forma dönüştürülmüştür. Eksik veya aşırı tekrar eden veri bulunmadığından veri temizleme süreci uygulanmamıştır. Ancak, etkisiz veya korelasyonu düşük olan bazı öznitelikler analiz dışında bırakılmış ve veri boyutu düşürülerek model doğruluğu artırılmıştır. Sonuç olarak, veri seti analizi 16 bağımsız değişken ve 1 hedef değişken ile gerçekleştirilmiştir.

Veri, model eğitiminde kullanılmak üzere %80 eğitim ve %20 test olacak şekilde train_test_split yöntemi ile ayrılmıştır.

Tablo 1. Çalışmada Kullanılan Değişkenler

Değişken Adı	Türü	Açıklama ve Değerler
Yaş	Nümerik	Değerler: 10'luk aralıklarla gruplanmıştır [0-10], [10-20], ..., [90-100]
Cinsiyet	Nominal	Değerler: Erkek, Kadın
Poliüri (sık idrara çıkma)	Nominal	Değerler: Yes / No
Polidipsi (aşırı susama)	Nominal	Değerler: Yes / No
Ani kilo kaybı	Nominal	Değerler: Yes / No
Yorgunluk	Nominal	Değerler: Yes / No
Bulanık görme	Nominal	Değerler: Yes / No
Kaşıntı	Nominal	Değerler: Yes / No
Yara iyileşmesinde gecikme	Nominal	Değerler: Yes / No
Vücuttaki tahriş	Nominal	Değerler: Yes / No
İnsülin kullanımı	Nominal	Değerler: Yes / No
Ailede diyabet geçmişi	Nominal	Değerler: Yes / No
Kaç kez su içme	Nominal	Değerler: Yes / No
Kaç kez idrara çıkma	Nominal	Değerler: Yes / No
Kaşıntı, bulanık görme, yorgunluk vb.	Nominal	Değerler: Yes / No
Açlık kan şekeri düzeyi	Nominal	Değerler: Yes / No
Obezite	Nominal	Değerler: Yes / No
İlaç kullanımı geçmişi	Nominal	Değerler: Yes / No
Sınıf (class)	Nominal	Diyabet durumu: Positive (diyabetli) / Negative (diyabetli değil)

2.2.Kullanılan Makine Öğrenmesi Algoritmaları

Çalışmada hedef değişken ‘class’ sütunu kullanılarak yedi denetimli algoritma eğitildi. Her algoritmanın eğitiminden önce nominal değişkenler “0” ve “1” formuna çevrilmiştir. (Yes:1, No:0, Male:1, Female:0, Positive:1, Negative:0). Sonraki aşamada hem sınıflandırma başarısı hem de yorumlanabilirlik açısından değerlendirilmiş, performans testi için şu metrikler kullanılmıştır; doğruluk, kesinlik, hassasiyet, F1 Skoru, karışıklık matrisi, ROC eğrisi, AUC değeri.

• Model Performans Değerlendirme Ölçütleri

Sınıflandırma modelinin performansını ölçütlerken karışıklık matrisindeki değerler, AUC değeri, F1 Skoru ölçütlerinin açıklamaları:

Doğruluk (Accuracy)

Modelin doğru tahmin ettiği tüm örneklerin toplam örneklere oranıdır. [10]

$$Doğruluk = \frac{DP + DN}{DP + DN + YP + YN}$$

Kesinlik (Precision)

Modelin pozitif olarak tahmin ettiği örneklerin ne kadarının gerçekten pozitif olduğunun oranıdır. [10]

$$Kesinlik = \frac{DP}{DP + YP}$$

Hassasiyet (Sensitivity)

Pozitif sınıfın doğru bir şekilde tanınma oranıdır ve gerçek pozitiflerin doğru tahmin edilme yüzdesini gösterir. [10]

$$Hassasiyet = \frac{DP}{DP + YN}$$

F1 Skoru (F1 Score)

Hassasiyet ve duyarlılığın harmonik ortalamasıdır ve dengesiz veri setlerinde modelin genel performansını ölçmek için kullanılır. [10]

$$F1 - Skoru = 2 \times \frac{Kesinlik * Hassasiyet}{Kesinlik + Hassasiyet}$$

AUC (Area Under Curve)

Modelin, farklı sınıfları doğru ayırt etme yeteneğini ölçen bir performans değeridir; ROC eğrisinin altındaki alanı ifade eder. [10]

Tablo 2. Karışıklık Matrisi

	Gerçek Değerler	
Tahmin edilen değerler	Positive	Negative
Positive	True Positive (TP)	False Positive (FP)
Negative	False Negative (FN)	True Negative (TN)

Bu kapsamda kullanılan yedi temel algoritma aşağıda açıklanmıştır:

2.2.1. Naive Bayes (Naif Bayes Sınıflandırıcısı)

Naive Bayes algoritması, olasılık temelli ve oldukça hızlı çalışan bir sınıflandırma yöntemidir. Bu model, her özelliğin sınıftan bağımsız olduğunu varsayarak Bayes teoremini uygular. Model uygulanırken “Early Stage Diabetes Risk Prediction” veri setinde yer alan kategorik ve sayısal değişkenler kullanılarak Gaussian Naive Bayes uygulanmıştır. [11] [13]

Çalışmamız özelinde:

- Algoritmanın doğruluk oranı %91 olarak tespit edilmiştir.
- Eğitim ve test verisi ayrımı yapıldıktan sonra model eğitilmiş ve sınıf tahmini yapılmıştır.

- Genellikle basit yapısı sayesinde hızlı sonuç vermiştir, ancak veri setindeki değişkenler bağımsız olmadığından diğer yöntemlere göre düşük doğruluk göstermiştir.

2.2.2. Logistic Regression (Lojistik Regresyon)

Lojistik regresyon, özellikle iki sınıflı (binary) problemler için yaygın olarak kullanılan doğrusal bir modeldir. İkili sınıflandırma amacıyla logit fonksiyonu içeren doğrusal sınır modeli tercih edildi. Logistic regresyon doğrusal ilişkiyi öğrenerek ilerler (Yaş arttıkça diyabet olma riskinin artması gibi) bu nedenle model sınıflar arasındaki sınırı doğrusal olarak çizmeye çalışır. [11] [13]

Çalışmamız özelinde:

- Bağımlı değişken olarak “class” etiketi (Positive / Negative) belirlenmiştir.
- Modelin doğruluk oranı oldukça yüksek olup, ROC eğrisi altında kalan alan (AUC) değeri 0,97 olarak gözlemlenmiştir.
- Değişkenlerin sınıflandırma üzerindeki etkisini ölçmek açısından yorumlanabilir sonuçlar üretmiştir.

2.2.3. Decision Tree (Karar Ağaçları)

Karar ağaçları, veriyi dallara ayırarak karar kuralları üreten ve görselleştirilmesi kolay olan bir yöntemdir. Ağaç yapısı içinde her düğüm bir özelliği ve eşik değerini temsil ederken, yapraklar sınıf etiketini içerir. Algoritma, her düğümde en iyi özelliği seçer ve bu özelliği kullanarak veriyi ikiye böler. Karar Ağacı, Gini İndeksi veya Entropi gibi ölçütlerle en iyi bölmeyi seçerek ilerler. [11] [13]

Çalışmamız özelinde:

- Model, eğitim verisi ile oluşturulmuş ve test verisi üzerinde doğrulama yapılmış ve başarı oranı 95% olarak bulunmuştur.
- Başarı oranı yükseltmek için etkisiz değişkenleri çıkartarak modeli tekrar eğitilmiştir.
- Etkisiz olarak “sudden weight loss” çıkartılmış, başarı oranının 98% olduğu görülmüştür ve overfitting olabilmeye riski ortaya çıkmıştır. Bunu overfitting olup olmadığını kontrol etmek için bazı değerlere (F1-skor, precision, recall, support) bakmamız gerekir.
- Bulgular ve Tartışma bölümündeki değerlere bakıldığında değerlerin birbirine oldukça yakın olduğu görülmüştür, bu da overfitting olmadığı sonucuna varılmasını sağlamıştır.
- ROC-AUC değeri ve doğruluk oranı yüksek çıkmıştır (AUC \approx 0.98).

2.2.4. Random Forest (Rastgele Orman)

Random Forest, birden fazla karar ağacının oluşturulup bunların sonuçlarının oylanarak sınıflandırma yapıldığı bir topluluk öğrenme yöntemidir. En son elde edilen tahminler, her bir ağacın verdiği kararların verdiği oylamasıyla birleştirilir, sonuç çoğunluk oylamasıyla belirlenir. Bu model hem overfitting’e karşı dayanıklı hem de yüksek doğruluklu sonuçlar verir. [11] [13]

Çalışmamız özelinde:

- 100'den fazla karar ağacı ile model eğitilmiş, her bir ağaç farklı örneklem veriler üzerinde çalıştırılmıştır.
- En iyi doğruluk ($\approx \%99$) ve AUC (1) değerine ulaşmıştır.
- Ulaşılan $\%99$ doğruluk ve AUC = 1, aşırı uyum (overfitting) şüphesi doğurdu. Bu yüzden bazı değerlere (F1-skor, precision, recall, support) bakmamız gerekti.
- Bu değerlere bakıldığında, Bulgular ve Tartışma bölümdeki Tablo 4 de görüldüğü gibi değerler birbirine oldukça yakın, bu da overfitting olmadığı kanıtına varılmasını sağlamıştır.

2.2.5. K-Nearest Neighbors (k-En Yakın Komşuluk)

Bu algoritma, test örneğini eğitim setindeki “k” adet en yakın veriyle karşılaştırarak sınıf tahmini yapar. KNN, mesafe ölçümü kullanarak çalışır, çoğunlukla Öklid mesafesi tercih edilir. Test verisi üzerinde tahmin yapmak için, eğitim setindeki “k” en yakın komşuyu bulur ve bu komşuların sınıflarına bakarak tahmin yapar. Özellikle örnekler arası benzerliğe dayalı basit ama etkili bir yöntemdir. [11] [13]

Çalışmamız özelinde:

- Modelde farklı “k” değerleri denenmiş ve en uygun $k=6$ değeriyle en iyi sonuç alınmıştır.
- Özellikler arasındaki uzaklıklar için öklidyen mesafe metriği kullanılmıştır.
- AUC ≈ 0.98 ve doğruluk oranı $\approx \%93.04$ olarak elde edilmiştir.

2.2.6. Linear Regression (Doğrusal Regresyon)

Linear Regression genellikle sürekli (sayısal) değerleri tahmin etmek için kullanılan bir modeldir. Ancak çalışılan veri seti ikili sınıflandırma problemine yönelik olduğundan uygun bir model değildir. Yine de numinal değerler binary değerlere çevrilerek tahmin edebilmesi sağlanmıştır. [11] [13]

Çalışmamız özelinde:

- Y_pred çıktısı round() fonksiyonu ile ikili sınıfa dönüştürülmüş ve confusion matrix üzerinden doğruluğu test edilmiştir.
- Sınıflandırma amacıyla kullanıldığında, doğrusal sınırlar üzerinden tahmin yapabildiği için bazı karmaşık örneklerde düşük performans göstermiştir.
- AUC ≈ 0.98 ve doğruluk oranı $\approx \%60.04$ olarak elde edilmiştir.
- Düşük oranı olduğu için eğitime devam edilmemiştir.

2.2.7. Yapay Sinir Ağı

Yapay sinir ağları (YSA), beyindeki nöron bağlantılarını matematiksel düğümlerle taklit ederek karmaşık ilişkileri öğrenebilen, denetimli bir öğrenme yöntemidir. Çok katmanlı algılayıcı (Multi-Layer Perceptron, MLP) mimarisi, her katmandaki nöronların ağırlıklı toplam-aktivasyon fonksiyonu mantığıyla çalışır; geriye yayılım (back-propagation) algoritması ise hatayı katmanlar arasında geriye doğru dağıtarak ağırlıkları günceller. [11] [13]

Deneyde kullanılan MLP'nin mimarisi şöyledir:

- Yöntem uygulanırken 16 giriş nöronu (sembiyotik semptom öznitelikleri), iki gizli katman (1. katman = 32 nöron, 2. katman = 16 nöron) ve tek çıkış nöronundan oluşan

bir MLP tasarlandı. Aktivasyon fonksiyonu gizli katmanlarda *tanh*, çıkış katmanında *sigmoid* olarak seçildi.

- Adam optimizasyonu (öğrenme hızı = 0.001) ve *binary cross-entropy* kayıp fonksiyonu ile 100 epoch'a kadar eğitim yapıldı; erken durdurma kriteriyle doğrulama kaybı 10 epoch boyunca iyileşmediğinde eğitim sonlandırıldı. Ve başarı oranı 98% bir model sonucu alınmıştır.

2.3.Literatür Taraması

Literatürde diyabet sınıflandırması üzerine yapılan birçok çalışmada Pima Indians Diabetes Dataset yaygın olarak kullanılmaktadır. Ancak bu veri seti yalnızca 21 yaş üzeri kadınlardan oluştuğu, verilerin eksik ve sınıfların dengesiz olması sebebiyle sonuçların genellenebilirliği tartışmalıdır. Ayrıca bazı makalelerde sadece sınırlı sayıda algoritma kullanılarak model karşılaştırması yapıldığı görülmüştür. Bizim araştırmamızda ise hem cinsiyet hem de semptom çeşitliliği bakımından daha kapsamlı olan Early Stage Diabetes veri seti tercih edilmiştir. Semptomlara dayalı olması nedeniyle düşük maliyetli tarama sistemleriyle entegrasyonu mümkündür. Ayrıca birden fazla algoritma kullanılarak performans analizi yapılmış ve hangi modelin daha iyi tahmin gücüne sahip olduğu karşılaştırmalı olarak değerlendirilmiştir. Bu yönüyle çalışma, literatürdeki benzer uygulamalardan ayrılmaktadır. [4] [8] [3]

Makalemizde kullanılan makine öğrenme algoritmaları, klasikleşmiş ve temel diyebileceğimiz yöntemlerdir. Veri setimizi kullanarak yapılan makalelerde ise daha karmaşık ve birleşik algoritmaların kullanılmasıyla başarı oranının önemli ölçüde arttığı gözlemlenmiştir. Bu, söz konusu veri setinin daha gelişmiş algoritmalarla işlendiğinde yüksek doğruluk oranlarına ulaşılabileceğini göstermektedir. Bununla birlikte, çalışmamızda Lineer Regresyon algoritması kullanılırken, özellikle karmaşıklık matrisinin çıkarılması ve veri setinin uygulanması aşamasında belirgin zorluklarla karşılaşmıştır. Bu durum, daha gelişmiş tekniklerin ve optimizasyon yöntemlerinin kullanımının önemini vurgulamaktadır. Bunlara ek olarak, semptom tabanlı sınıflandırmada doğrusal olmayan etkileşimleri yakalayabilen YSA'nın klasik algoritmalarla kıyasla daha az öznitelik mühendisliği gerektirdiği gözlemlendi. Pima Indians veri setiyle yapılan önceki YSA çalışmalarında doğruluk değerleri % 76–94 aralığında raporlanmıştır [12] [9]

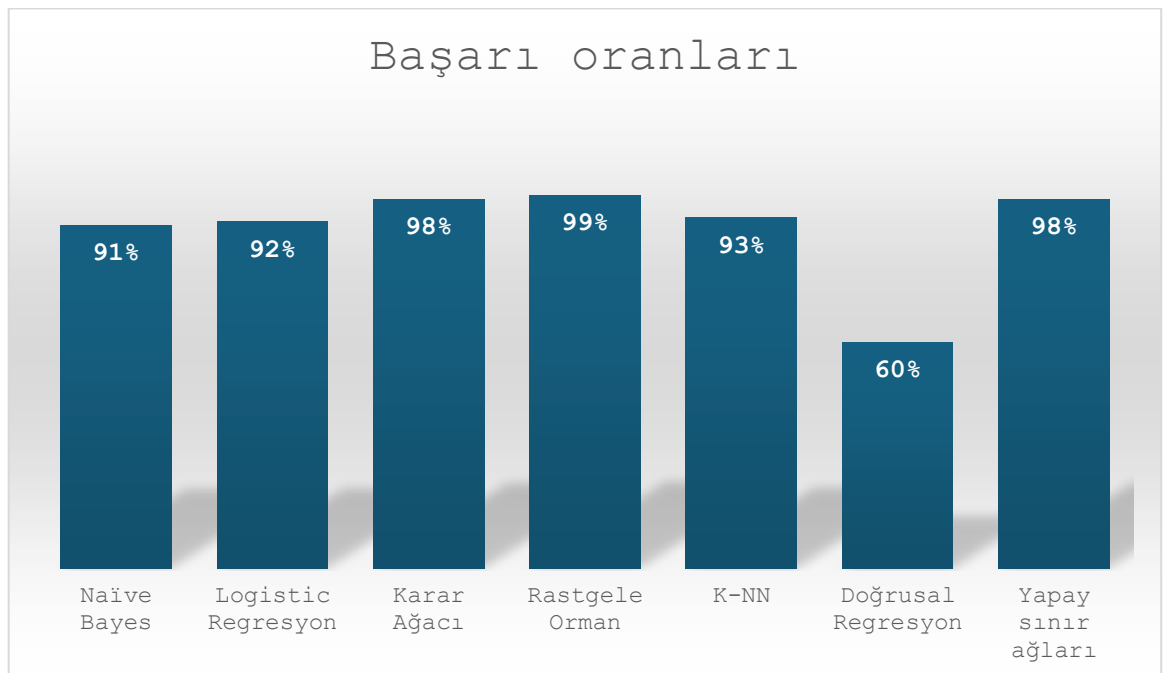
Makaleyi oluştururken karşılaştığımız zorluklardan bahsedecek olursak, veri setimizin oldukça temiz olması nedeniyle bazı algoritmaların beklediğimizden daha başarılı sonuçlar vermesi karşılaştığımız ilk zorluktu. Bu durum, hangi algoritmanın overfitting yapıp yapmadığını tespit etmek ve kontrolünü sağlamak noktasında zorluklar yaşamamıza sebep oldu. Sonuçları değerlendirirken her bir skoru tek tek gözden geçirmemiz gerekti ve bu süreçte her algoritma için tekrarlar yapmamız gerektiği için bu durum oldukça yorucu oldu. Kullanılan beş algoritmanın her birinde overfitting riskiyle karşılaştık. Özellikle Lineer Regresyon algoritmasında başarı oranı diğer algoritmalarla kıyaslandığında çok düşük kaldığı için underfitting ihtimalini göz önünde bulundurduk. Ancak yaptığımız literatür taramasında, lineer regresyonun sürekli değişkenler için tercih edilen bir yöntem olduğunu ve bu nedenle başarı oranının daha düşük olduğunu öğrendik. Bu bilgi, modelin düşük başarı oranını açıklamak adına önemli bir bulgu oldu.

3. Bulgular ve Tartışma

Modeller, semptom değişkenleri üzerinde beş metrikle karşılaştırıldı ve sonuçlar Tablo 4'te özetlendi. Modellerin başarısını ölçümlemek için kullanılan doğruluk oranı, AUC, hassasiyet ve karışıklık matrisi gibi ölçütler, her bir modelin veri setindeki sınıflandırma doğruluğunu belirlemede etkili olmuştur.

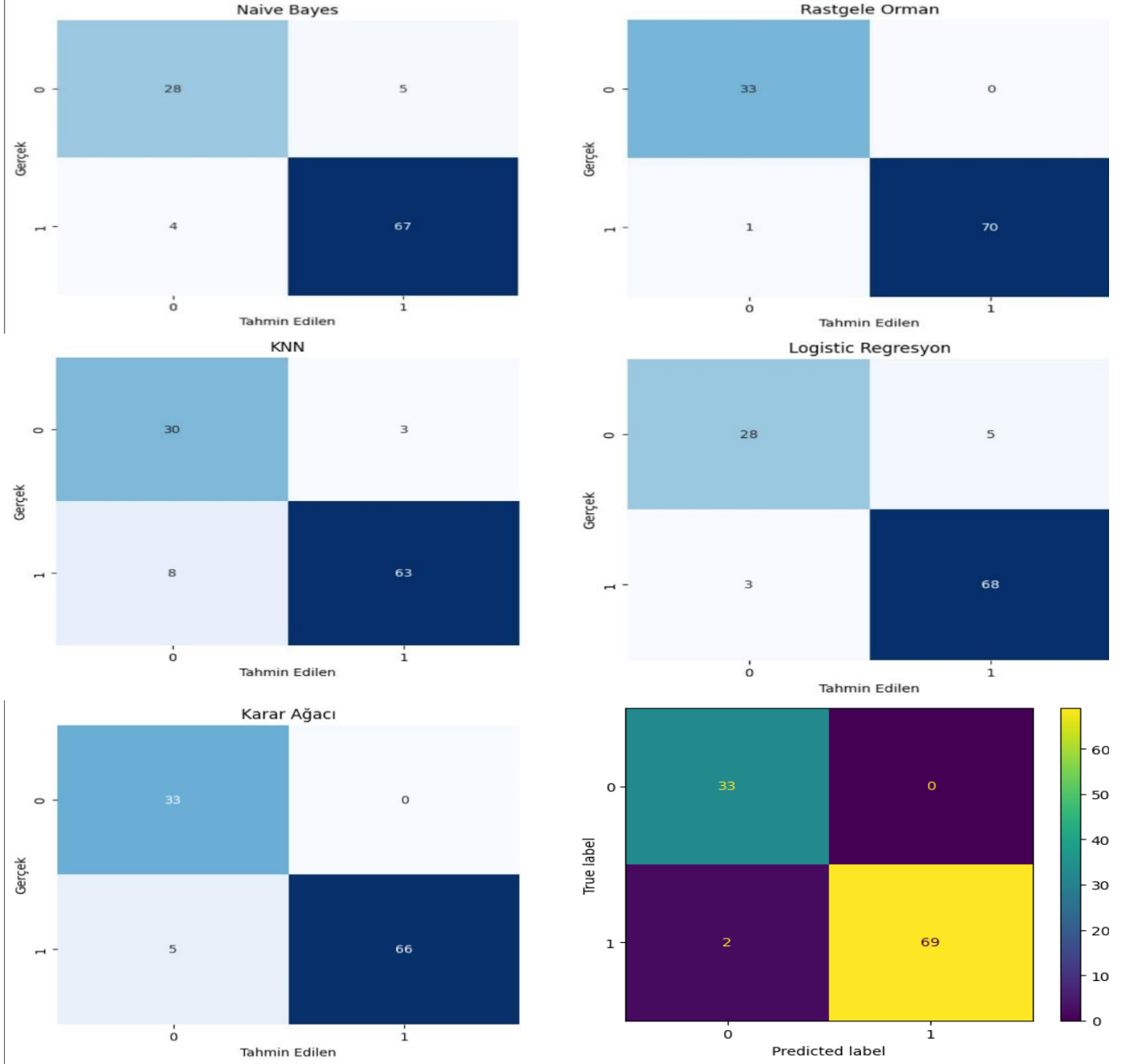
- **Karar Ağacı (Decision Tree)** modeli, %98 doğruluk oranı ile en yüksek 3. model olmuştur. Bu model, veri setindeki önemli özellikleri doğru bir şekilde ayırıştırarak, sınıflandırma görevini başarıyla tamamlamıştır.
- **Rastgele Orman (Random Forest)**, %99 başarı oranı ile en iyi performansı göstermiştir. Rastgele Orman, birçok karar ağacının birleşimiyle güçlü bir sınıflandırma modeli oluşturmuş ve yüksek doğruluk sağlamıştır.
- **K-En Yakın Komşu (K-NN)** modeli, %93 doğruluk oranı ile yüksek bir başarı elde etmiştir. Ancak, KNN, daha fazla hesaplama gücü gerektiren bir algoritma olup, doğruluk oranı diğer modellerin gerisinde kalmıştır.
- **Lojistik Regresyon (Logistic Regression)**, %92 doğruluk oranı ile başarılı bir model olmasına rağmen, daha karmaşık modeller kadar yüksek doğruluk elde edememiştir.
- Naive Bayes, %91 doğruluk oranı ile uygun sonuçlar elde etmiştir ancak genellikle diğer modellerin gerisinde kalmıştır. Özellikle bağımsızlık varsayımının bazı özellikler için geçerli olmaması, modelin başarısını sınırlamıştır.
- **Doğrusal Regresyon (Linear Regression)**, %60 doğruluk oranı ile en düşük başarıyı sağlamıştır. Bu modelin doğrusal sınıflandırma yapması, daha karmaşık veri setlerinde doğru tahmin yapabilmesi için yeterli olmamıştır.
- **Yapay Sinir Ağı (Artificial Neural Network, YSA)** modeli, %97 doğruluk oranı ile Rastgele Orman'ın hemen ardından en yüksek 2. başarıyı sergilemiştir. Modelin eğitimi uzun sürmüş, aktivasyon (tanh ↔ ReLU) ve katman başına nöron sayısı gibi hiperparametrelere duyarlılığı nedeniyle dikkatli optimizasyon gerektirmiştir

Aşağıda, bu modellerin başarı oranlarını daha net bir şekilde görselleştiren grafiği bulabilirsiniz:



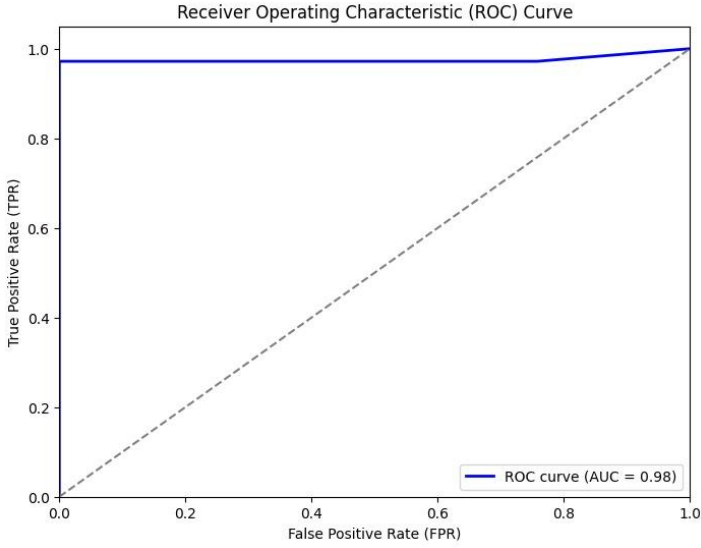
Makine Öğrenmesi ile Erken Evre Diyabet Riskinin Sınıflandırılması

Tablo 3. Modellerin Karışıklık Matrisleri

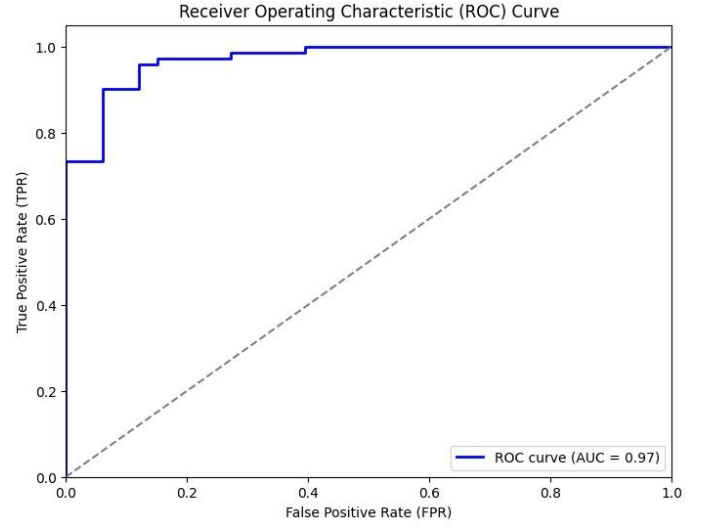


- **Naive Bayes** modeli, 28 doğru negatif (True Negative) ve 67 doğru pozitif (True Positive) tahmin ile iyi bir performans sergileyerek, sınıflandırma başarısını göstermektedir. **Rastgele Orman** algoritması, 33 doğru negatif ve 70 doğru pozitif tahminle en yüksek doğruluk oranını elde etmiş olup, bu modelin doğruluğu diğer modellere göre daha yüksektir. **KNN** modelinde ise 30 doğru negatif ve 63 doğru pozitif tahmin yapılmış, bu da modelin kabul edilebilir bir doğruluk seviyesine sahip olduğunu göstermektedir. **Logistic Regresyon** modelinde 28 doğru negatif ve 68 doğru pozitif tahmin gerçekleştirilmiştir, bu da modelin başarılı olduğunu ancak diğer bazı modeller kadar yüksek performans sergilemediğini göstermektedir. **Karar Ağacı** modeli ise 33 doğru negatif ve 66 doğru pozitif tahminle, doğruluk açısından tatmin edici bir sonuç sunmaktadır. **Yapay Sinir Ağı (YSA)** modelinde de 33 doğru negatif ve 69 doğru pozitif tahmin elde edilmiş, bu performansı ile Rastgele Orman modelinin hemen ardından gelerek yüksek bir doğruluk seviyesi ortaya koymuştur. Bu sonuçlar, her bir modelin sınıflandırma doğruluğunu yansıtarak, hangi algoritmanın daha verimli olduğunu belirlemeye olanak sağlamaktadır.

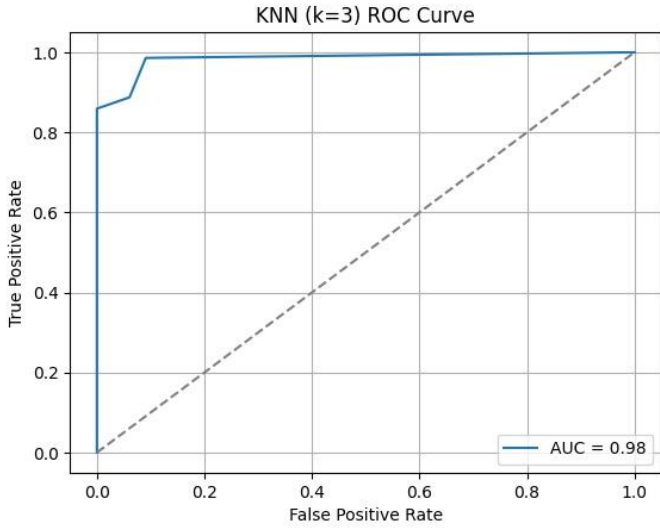
Makine Öğrenmesi ile Erken Evre Diyabet Riskinin Sınıflandırılması



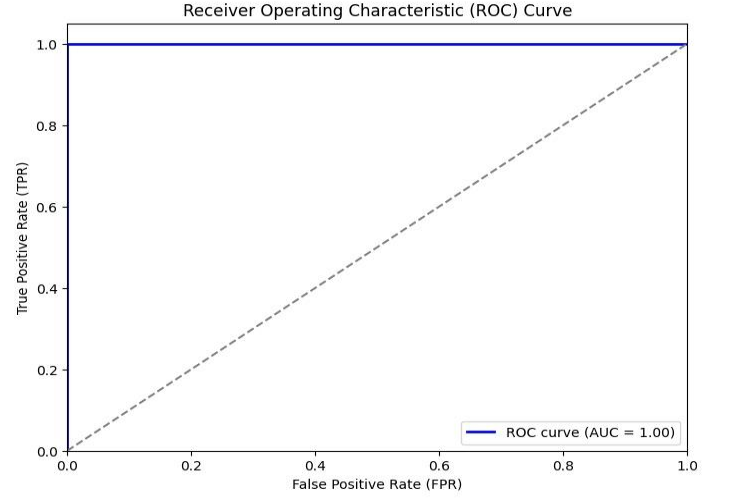
ii. Karar Ağacı



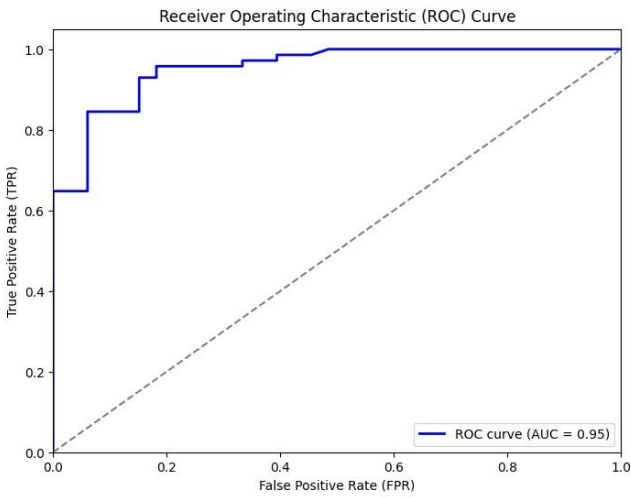
iii. Logistic Regresyon



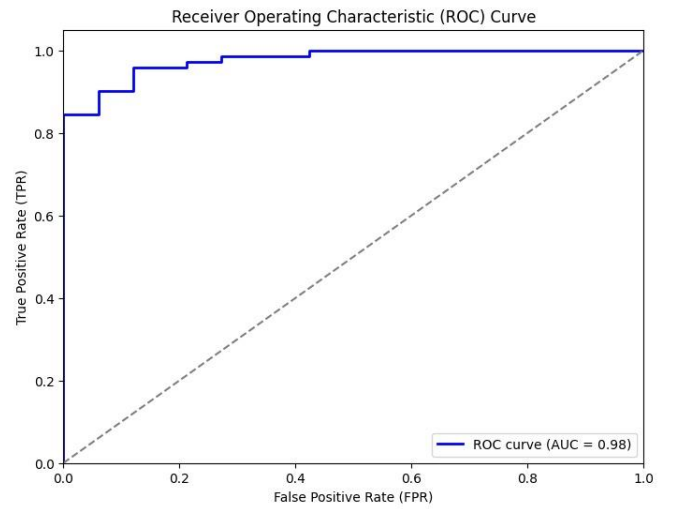
iv. K-NN



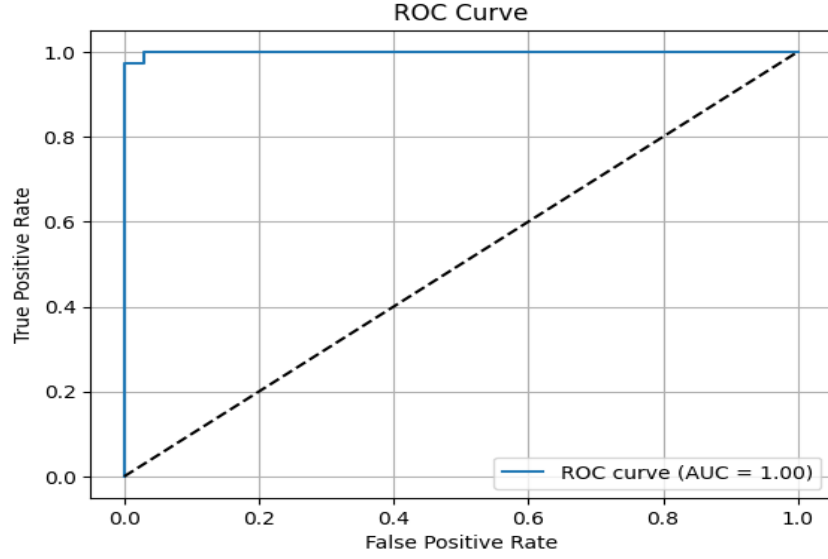
v. Rastgele Orman



vi. Naive Bayes



vii. Lineer Regresyon

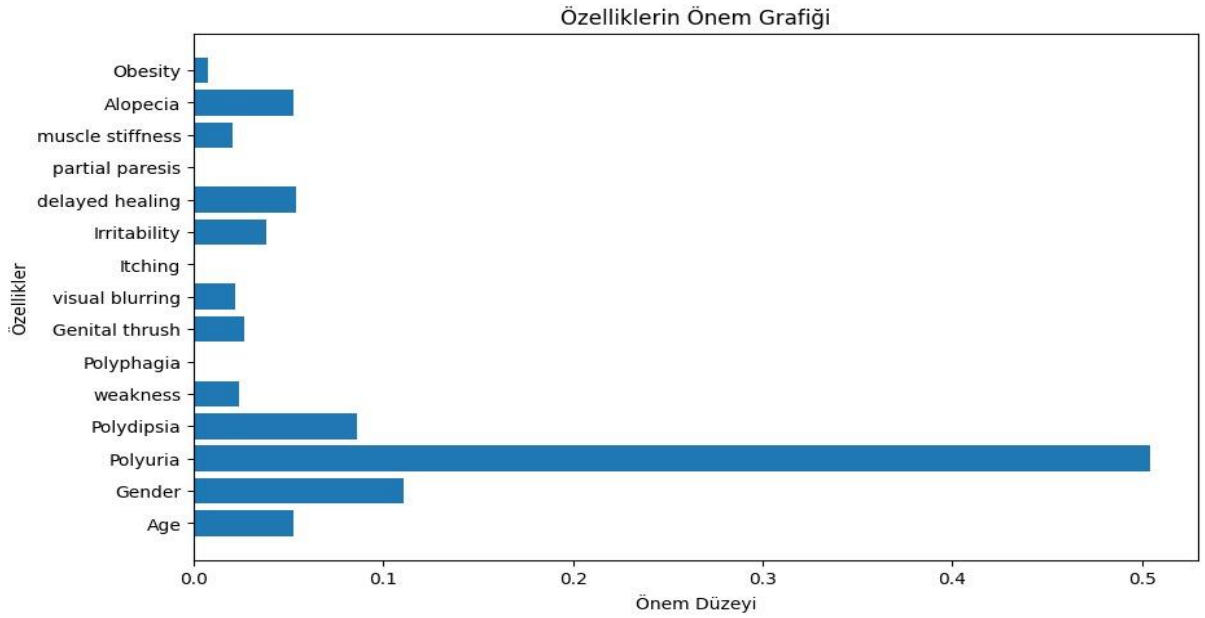


viii. Yapay Sinir Ağı

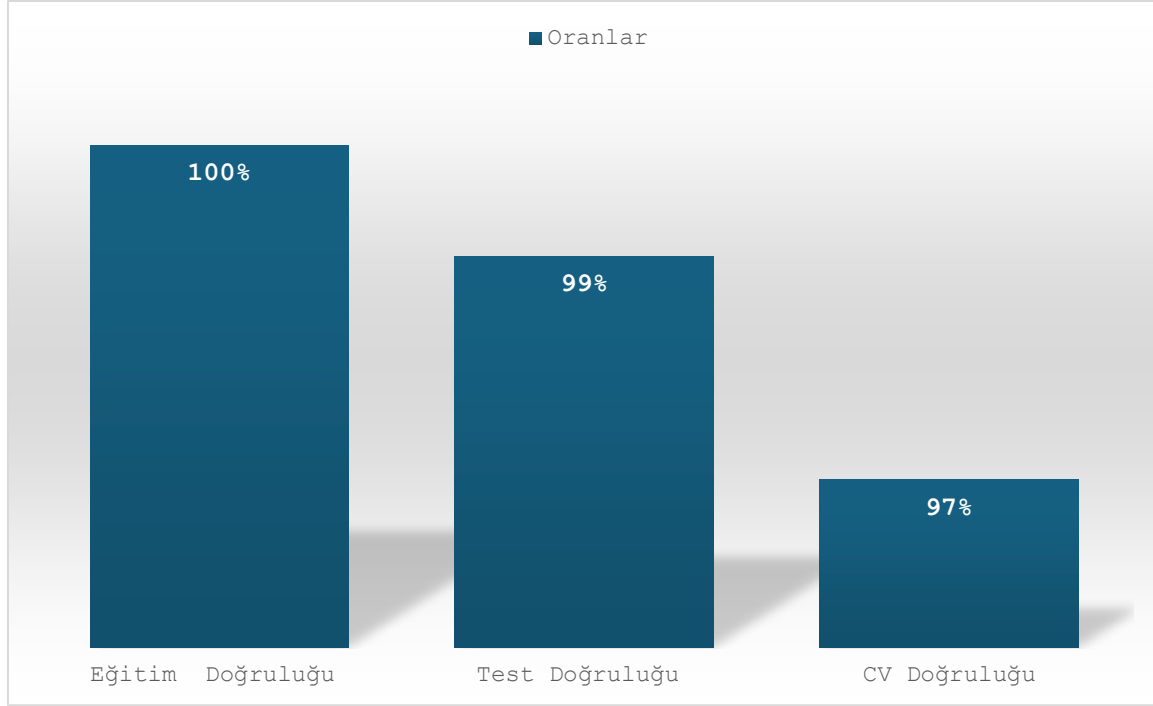
- AUC (Area Under the Curve) ve ROC (Receiver Operating Characteristic) eğrisi, sınıflandırma modelinin performansını değerlendirmek için kullanılan önemli iki metriği ifade eder. ROC eğrisi, modelin doğruluk oranını gösterirken, False Positive Rate (FPR) ile True Positive Rate (TPR) arasındaki ilişkiyi görsel olarak sunar. AUC değeri ise bu eğrinin altındaki alanı temsil eder ve modelin genel doğruluğu hakkında bilgi verir. AUC değeri 0 ile 1 arasında değişir; 1'e yakın değerler modelin mükemmel performansını, 0.5'e yakın değerler ise rastgele tahminle yakın bir başarıyı gösterir. Örneğin, **Karar Ağacı** modeli AUC değeri 0.98 ile oldukça yüksek bir başarıya sahipken, **Logistic Regresyon** modeli 0.97 AUC ile yine yüksek bir performans sunmaktadır. **KNN** modeli, AUC 0.98 ile başarılı sonuçlar verirken, **Rastgele Orman** modeli AUC değeri 1.00 ile mükemmel sonuçlar elde etmiştir, bu da modelin neredeyse kusursuz olduğunu gösterir. **Naive Bayes** ve **Linear Regresyon** modelleri ise sırasıyla AUC 0.93 ve 0.90 ile iyi performans sergilemiş olsa da diğer modellere göre biraz daha düşük performans göstermektedir. **Yapay Sinir Ağı (YSA)** modeli de ROC eğrisinden hesaplanan AUC değeri 1.00 ile Rastgele Orman ile birlikte en yüksek performansı sergileyerek sınıflandırma görevinde kusursuz bir ayırım gücü ortaya koymuştur. ROC eğrisi, modelin doğru sınıflandırma oranını, yanlış sınıflandırma oranıyla karşılaştırarak modelin tahmin gücünü net bir şekilde görmemize olanak tanır.

Tablo 4. Performans Ölçütleri

Model	Kesinlik	Doğruluk	Hassasiyet	AUC	F1-Skoru
Naive Bayes	0.90	0.91	0.90	0.95	0.90
K-NN	0.87	0.93	0.90	0.98	0.88
Lojistik Regresyon	0.92	0.92	0.90	0.97	0.91
Karar Ağaçları	0.97	0.98	0.99	0.98	0.98
Rastgele Orman	0.99	0.99	0.99	1.00	0.99
Lineer Regresyon	0.92	0.60	0.90	0.98	0.91
Yapay Sinir Ağı	0.96	0.97	0.98	1.00	0.97



Tablo 5. Karar Ağacı için Değişkenlerin Etkinlikleri



Tablo 6. Rastgele Orman için Overfitting risk kontrolü

4. Etik Beyan

Bu çalışma, UCI Machine Learning Repository’de kamuya açık ve kimliksizleştirilmiş *Early Stage Diabetes Risk Prediction* veri kümesini kullanmıştır. Kişisel veya tanımlayıcı bilgi içermediğinden etik kurul onayı gerektirmemektedir.

5. Sonuç ve Öneriler

Erken diyabet taraması için toplam yedi algoritma test edilmiştir. Yaptığımız deneyler sonucunda en yüksek başarıyı *Rastgele Orman* algoritmasında elde ettik, *Yapay Sinir Ağı* modeli ise %97 doğruluk ile ikinci sırada yer almıştır. Rastgele Orman modeli, yüksek doğruluk oranıyla diyabetin erken evrelerini doğru bir şekilde sınıflandırmada etkili olmuştur.

Diyabetin erken evrede tespiti ve tedavisine erken başlanması, ileri yaşamda oluşabilecek sıkıntıları önlemek için en önemli noktadır ve bu bizim çalışmamızın temel amacını oluşturur. Bu noktada, makine öğrenmesi teknolojilerinin kullanımı oldukça kritik bir rol oynamaktadır. Çünkü geleneksel yöntemler, verileri işlemek ve anlamlı sonuçlara ulaşmak için yeterli olmayabiliyor. Makine öğrenmesi algoritmaları, verilerin karmaşıklığını daha etkili bir şekilde analiz edebilir ve doğru sonuçlar üretebilir.

Gelecek araştırmalarda, daha gelişmiş algoritmaların kullanılması önerilmektedir. Özellikle derin öğrenme (mevcut araştırmada test edilen MLP dâhil) ve XGBoost gibi güçlü modellerin de araştırmaya dahil edilmesi, diyabet riski tahmininde daha doğru sonuçlar elde edilmesini sağlayabilir. Ayrıca, günümüzde diyabetin genç yaşlarda daha sık görülmeye başlaması, veri toplama sürecine yeni bir boyut kazandırmaktadır. Kullanılacak veri seti için genç bireylerden de veri toplanması, cinsiyet oranının daha dengeli olması gibi faktörler, veri setinin daha esnek ve çeşitli olmasına yardımcı olacaktır. Bu değişiklikler de modelin genel geçerliliğini artırmasına ve daha geniş bir popülasyona hitap etmesine sebep olacaktır. Bu araştırmanın bulguları, diyabetin erken evre tanısı için makine öğrenmesi teknolojilerinin etkinliğini bir kez daha gözler önüne sermektedir. İleride yapılacak olan çalışmalarda bu alanda yeni algoritmalar ve daha geniş veri setleri kullanarak daha kapsamlı sonuçlar üretmesi beklenmektedir. [5] [9] [2] [13]

Kaynakça:

- [1] Centers for Disease Control and Prevention, «Diabetes,» 2022. [Çevrimiçi]. Available: <https://www.cdc.gov/diabetes/basics/diabetes.html>.
- [2] World Health Organization, «Diabetes,» 2023. [Çevrimiçi]. Available: <https://www.who.int/news-room/fact-sheets/detail/diabetes>.
- [3] Ö. N. Ergün ve O. İlhami, «Early Stage Diabetes Prediction Using Machine Learning Methods,» *Avrupa Bilim ve Teknoloji Dergisi*, cilt 29, 2021.
- [4] S. Firdous, G. A. Wagai ve K. Sharma, «A survey on diabetes risk prediction using machine learning approaches,» *Journal of Family Medicine and Primary Care*, cilt 11, no. 11, pp. 6929-6934 , 2022.
- [5] International Diabetes Federation, «IDF Diabetes Atlas (10th ed.),» International Diabetes Federation, Brüksel, 2021.
- [6] T.C. Sağlık Bakanlığı, «2018 Bütçe Sunumu,» T.C. Sağlık Bakanlığı, Ankara, 2018.
- [7] Türkiye Diyabet Vakfı, «2016-2017 Raporları,» 2017. [Çevrimiçi]. Available: <https://www.turkdiab.org/>.
- [8] Kaggle, «Pima Indians Diabetes Database,» 2020. [Çevrimiçi]. Available: <https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database>.
- [9] UCI Machine Learning Repository, «Early Stage Diabetes Risk Prediction Dataset,» 2019. [Çevrimiçi]. Available: <https://archive.ics.uci.edu/dataset/529/early+stage+diabetes+risk+prediction+dataset>.
- [10] A. Çınar, «Veri Madenciliğinde Sınıflandırma Algoritmalarının Performans Değerlendirmesi ve R Dili ile Bir Uygulama,» *Öneri Dergisi*, cilt 14, no. 51, pp. 90-111 , 2019.
- [11] O. Seveli, «Göğüs Kanseri Teşhisinde Farklı Makine Öğrenmesi Tekniklerinin Performans Karşılaştırması,» *Avrupa Bilim ve Teknoloji Dergisi*, no. 16, pp. 179-185, 2019.
- [12] International Journal of Computer Applications, «Volume 45 – No. 12 (May 2012),» 2012. [Çevrimiçi]. Available: <https://www.ijcaonline.org/archives/volume45/number12/6836-9460/>.
- [13] M. Y. ., E. S. S. Bilge Özlüer Başer, «Makine Öğrenmesi Teknikleriyle Diyabet Hastalığının Sınıflandırılması,» *Süleyman Demirel Üniversitesi Fen Bilimleri Enstitüsü Dergisi*, cilt 25, pp. 112-120 , 2021.
- [14] <https://github.com/21040001/Early-Stage-Diabetes-Risk-Prediction->