



# MAKİNE ÖĞRENMESİ MODELLERİNİN DİYABET TAHMİNİNDEKİ PERFORMANSLARININ KARŞILAŞTIRILMASI

Yazılım Mühendisliği Ana Bilim Dalı

Yüksek Lisans Bitirme Projesi

**Recai Cem KAHVECİ**

Proje Danışmanı: Doç. Dr. Aytuğ ONAN

Ocak 2024

Proje Başlığı: “Makine Öğrenmesi Modellerinin Diyabet Tahminindeki Performanslarının Karşılaştırılması”

## Öz

Bu çalışma, makine öğrenmesi algoritmalarının kullanıldığı bir diyabet tahmin modelinin geliştirilmesi ve performansının karşılaştırılması üzerine odaklanmaktadır. Veri seti, "diabetes.csv" adlı dosyada bulunan klinik ve demografik özelliklere dayanmaktadır. Çalışma, model performansını değerlendirmek için doğruluk, hassasiyet, özgüllük, F1 puanı, ROC eğrisi ve AUC değeri gibi çeşitli değerlendirme ölçütlerini içermektedir.

Çalışmanın ilk aşamasında, veri seti keşfedilmiş, temizlenmiş ve ön işleme adımları gerçekleştirilmiştir. Daha sonra, lojistik regresyon, Naive Bayes, KNN, SVM, karar ağaçları, random forests, gradient boosting, XGBoost, CatBoost ve yapay sinir ağları gibi farklı makine öğrenmesi algoritmaları kullanılarak modeller oluşturulmuştur.

Elde edilen modellerin performansı çeşitli metriklerle değerlendirilmiştir. Doğruluk, hassasiyet, özgüllük, F1 puanı ve AUC değeri gibi ölçütler kullanılarak modellerin karşılaştırılması yapılmıştır. Sonuçlar, her bir modelin diyabet tahminindeki gücünü ve zayıflıklarını ortaya koymaktadır.

Bu çalışma, diyabet tahmininde kullanılan farklı makine öğrenmesi modellerinin performansını değerlendirerek, hangi modelin daha etkili olduğunu belirlemeyi amaçlamaktadır.

Anahtar Sözcükler: Makine öğrenmesi, mühendislik, yapay zeka, klinik uygulamalar, biyomedikal veri analizi

Project Title: "Comparison of the Performance of Machine Learning Models in Diabetes Prediction"

## **Abstract**

This study focuses on the development and comparison of a diabetes prediction model using machine learning algorithms. The dataset relies on clinical and demographic features found in the "diabetes.csv" file. The evaluation of model performance includes various metrics such as accuracy, precision, specificity, F1 score, ROC curve, and AUC value.

In the initial stage of the study, the dataset was explored, cleaned, and subjected to preprocessing steps. Subsequently, different machine learning algorithms, including logistic regression, Naive Bayes, KNN, SVM, decision trees, random forests, gradient boosting, XGBoost, CatBoost, and artificial neural networks, were used to create models.

The performance of the obtained models was assessed using various metrics. Models were compared based on criteria such as accuracy, precision, specificity, F1 score, and AUC value. The results reveal the strengths and weaknesses of each model in diabetes prediction.

This study aims to determine the effectiveness of different machine learning models used in diabetes prediction by evaluating their performance.

Keywords: Machine learning, engineering, artificial intelligence, clinical applications, biomedical data analysis

## Teşekkür

Tüm eğitim süreci boyunca, öğrenciler olarak karşılaştığımız her zorlukta bizlere sonuna kadar destek olan ve bu proje çalışmamda kendisi ile çalışma isteğimi kabul ettiği için Sayın

Doç. Dr. Aytuğ ONAN'a teşekkürlerimle...

# İçindekiler

Öz	i
Abstract	ii
Teşekkür	iii
1. Giriş	1
2. Kullanılan Modeller	2
2.1. Lojistik Regresyon	2
2.2. Naive Bayes	2
2.3. K-En Yakın Komşu	2
2.4. Destek Vektör Makineleri	3
2.5. Karar Ağaçları	3
2.6. Rastgele Ormanlar	3
2.7. Gradient Boosting	4
2.8. XGBoost	4
2.9. CatBoost	4
2.10. Yapay Sinir Ağları	5

3. Kullanılan Yöntem	5
3.1. Gerekli Bağımlılıkların Projeye Eklenmesi	5
3.2. Veri Setinin Yüklenmesi	6
3.3. Veri Setinin İncelemesi	6
3.4. Bağımsız ve Bağımlı Değişkenlerin Belirlenmesi	7
3.5. Veri Setinin Eğitim ve Test Setlerine Bölünmesi	8
3.6. Modelin Oluşturulması ve Eğitilmesi	8
3.7. Tahmin Yapılması ve Performans Metriklerinin Hesaplanması	8
3.8. ROC Eğrisi ve AUC Değerinin Çizilmesi	9
3.9. Grafikselleştirme ile Sonuçların İfade Edilmesi	10
4. Elde Edilen Sonuçların Değerlendirilmesi	11
4.1. Bulgular	11
4.2. Bulguların Yorumlanması ve Sonuç	21
Kaynaklar	23

## Şekiller Listesi

Şekil 1: Lojistik Regresyon Çalışması İçin Import Edilen Bağımlılıklar	5
Şekil 2: Veri Setinin Yüklenmesi	6
Şekil 3: Veri Setinin İlk Beş Satırının İncelenmesi	6
Şekil 4: Veri Setinin İstatistiklerinin İncelenmesi	7
Şekil 5: Bağımlı ve Bağımsız Değişkenlerin Belirlenmesi	7
Şekil 6: Verinin Eğitim ve Test Setlerine Bölünmesi	8
Şekil 7: Extreme Gradient Boosting Machines Modelinin Oluşturulması ve Eğitilmesi	8
Şekil 8: Modeli Kullanarak Tahmin Yapılması	8
Şekil 9: Performans Metriklerinin Ölçülmesi	9
Şekil 10: ROC Eğrisinin ve AUC Değerinin Çizilmesi	9
Şekil 11: ROC Eğrisinin Grafikselleştirilmesi	10
Şekil 12: Karar Ağaçları Modelinin Performansı	11
Şekil 13: Naive Bayes Modelinin Performansı	12
Şekil 14: K-En Yakın Komşu Modelinin Performansı	13
Şekil 15: Destek Vektör Makineleri Modelinin Performansı	14
Şekil 16: Yapay Sinir Ağları Modelinin Performansı	15
Şekil 17: CatBoost Modelinin Performansı	16
Şekil 18: Karar Ağaçları Modelinin Performansı	17

Şekil 19: Random Forests Modelinin Performansı	18
Şekil 20: Gradient Boosting Machines Modelinin Performansı	19
Şekil 21: Extreme Gradient Boosting Machines (XGBoost) Modelinin Performansı	20
Şekil 22: Tüm Performans Metriklerinin Bir Arada Gösterimi	21
Şekil 23: Tüm Modellerin AUC Skorları (Sıralı)	22



## 1. Giriş

Makine öğrenmesi alanındaki hızlı gelişmeler, sağlık sektöründe kullanılan tahmin modellerinin etkinliğini artırmak için yeni olanaklar sunmaktadır. Bu bağlamda, bu çalışma, diyabet tahmininde kullanılan farklı makine öğrenmesi modellerinin performansını karşılaştırmayı amaçlamaktadır.

Çalışmanın temel amacı, klinik ve demografik özelliklere dayalı bir veri seti üzerinde çeşitli makine öğrenmesi algoritmaları kullanılarak oluşturulan modellerin diyabet tahminindeki güçlerini ve zayıflıklarını değerlendirmektir.

Bu kapsamda, lojistik regresyon, Naive Bayes, KNN, SVM, karar ağaçları, random forests, gradient boosting, XGBoost, CatBoost ve yapay sinir ağları gibi çeşitli makine öğrenmesi modelleri kullanılmıştır.

Modeller, çeşitli değerlendirme ölçütleri kullanılarak kapsamlı bir performans analizi için karşılaştırılmıştır. Doğruluk, hassasiyet, özgüllük, F1 puanı, ROC eğrisi ve AUC değeri gibi metrikler, her bir modelin diyabet tahminindeki başarı düzeyini belirlemede temel kriterler olarak ele alınmıştır.

Bu çalışmanın önemi, diyabetin erken teşhisi ve etkili bir şekilde yönetilmesi için kullanılacak bir modelin seçimine yönelik bilgi sağlamaktadır.

Kullanılan modellerin performansının ayrıntılı bir şekilde değerlendirilmesi, sağlık profesyonellerine ve araştırmacılara daha iyi anlayış ve güvenilir bir araç sunma potansiyeline sahiptir.

İlerleyen bölümlerde, her bir modelin detaylı açıklamalarını sunarak, elde edilen sonuçların daha kapsamlı bir şekilde anlaşılmasına odaklanılacaktır.

## 2. Kullanılan Modeller

### 2.1. Lojistik Regresyon

Çalışmada kullanılan ilk model, lojistik regresyon, diyabet tahmininde sıklıkla kullanılan bir yöntemdir. Bu model, bağımlı değişkenin olasılık dağılımını modeller ve veri setindeki bağımsız değişkenlerle ilişki kurar (Ganesh vd., 2021).

Lojistik regresyon, basit yapıya sahip olmasının yanı sıra yüksek derecede yorumlanabilirlik sunar, bu da sağlık profesyonellerinin modelin tahminlerini anlamalarını kolaylaştırır (Zhu vd., 2019).

### 2.2. Naive Bayes

İkinci olarak, Naive Bayes modeli kullanılmıştır. Bu model, Bayes teoremi temel alınarak çalışır ve bağımsız değişkenler arasındaki koşullu bağımlılık varsayımı yapar (Hasanah & Munggaran, 2021). Naive Bayes, özellikle sınıflandırma çalışmalarında etkili olduğu için diyabet tahmini için uygun bir modeldir (Okikiola vd., 2023). Model, basit yapısı ve hızlı eğitim süreleri ile bilinir (Prakash vd., 2021).

### 2.3. K-En Yakın Komşu

KNN (K-Nearest Neighbors) modeli de çalışmada kullanılan bir diğer yaklaşımdır. Bu model, benzer özelliklere sahip olan veri noktalarını gruplandırarak, bir noktanın sınıfını belirler (Sinaga et al., 2020).

KNN'nin esnek yapısı ve parametre ayarı kolaylığı, diyabet veri setindeki karmaşıklıklarla başa çıkmak için kullanışlıdır (Gupta & Goel, 2020).

## 2.4. Destek Vektör Makineleri

SVM (Support Vector Machines), lineer ve non-lineer problemlerde etkili olan bir modeldir. Özellikle veri setindeki karmaşıklıkları çözmek için kullanılan bu model, maksimum marjinal ayırım prensibi üzerine kuruludur ve özellikle yüksek boyutlu veri setlerinde performans gösterir (Jain, 2022). Diyabet tahmini çalışmalarında daha önce de kullanılmış bir modeldir (Joshi & Chawan, 2018).

## 2.5. Karar Ağaçları

Karar Ağaçları, veri setindeki desenleri basit karar kurallarıyla açıklamak ve tahminlerde bulunmak için kullanılan bir diğer modeldir. Bu model, veri setini özelliklere göre bölerek ve her bir bölme noktasında kararlar alarak bir ağaç yapısı oluşturur. Her yaprak düğümü, belirli bir sınıfa ait veri noktalarını temsil eder (Hasanah & Munggaran, 2021).

Karar Ağaçları, modelin diyabet verisinde yüksek derecede yorumlanabilir ve açıklanabilir olmasını sağlar (Jain, 2022).

## 2.6. Rastgele Ormanlar

Random Forests, bu çalışmada kullanılan bir başka önemli modeldir. Bu model, birden çok karar ağacını birleştirerek, daha güçlü ve stabil bir tahmin modeli oluşturmayı amaçlar. Her bir ağaç, rastgele seçilen alt özellikler üzerinde eğitilir, bu da overfitting'i azaltır ve genelleme yeteneğini artırır. (Bonfietti vd., 2015)

Random Forests, veri setindeki gürültüye ve karmaşıklığa karşı dayanıklıdır, bu da onu geniş bir uygulama yelpazesinde etkili kılar (Baher vd., 2020).

## 2.7. Gradient Boosting

Gradient Boosting, zayıf öğrenicileri bir araya getirerek güçlü bir tahmin modeli oluşturan bir ensemble yöntemidir. Bu model, ardışık olarak oluşturulan zayıf öğrenicilerin, hataları üzerine odaklanarak, daha iyi bir tahmin yapma yeteneğine sahiptir. Her bir önceki modelin hatalarını düzeltmeye odaklanan Gradient Boosting, karmaşık ilişkileri modelleme yeteneğiyle öne çıkar (Solichin vd., 2019).

## 2.8. XGBoost

XGBoost, eXtreme Gradient Boosting'in kısaltmasıdır ve Gradient Boosting'in bir türevidir. XGBoost, özellikle hız ve performans üzerine odaklanarak, büyük veri setlerinde etkili bir şekilde çalışabilen bir modeldir. Regularizasyon teknikleri, ağaç tabanlı öğreniciyi güçlendiren bu modelin, overfitting'e karşı dayanıklılığını artırır (Marzi vd., 2023).

XGBoost, genellikle rekabetçi veri bilimi yarışmalarında ve endüstriyel uygulamalarda tercih edilen bir modeldir (Kapoor & Perrone, 2021). Sağlık alanında daha önce farklı hastalıkların tahmini için başarılı sonuçlar vermiş bir modeldir (Liu vd., 2022).

## 2.9. CatBoost

CatBoost, kategorik değişkenleri doğrudan işleyebilen bir gradient boosting algoritmasıdır. Bu özellik, diğer geleneksel gradient boosting modellerinde karşılaşılan kategorik değişkenlerle başa çıkma zorluğunu aşmada CatBoost'u öne çıkarır. Model, kategorik değişkenlerin içerdiği bilgiyi daha etkili bir şekilde kullanarak, tahmin performansını artırır (Dorogush vd., 2018).

CatBoost, ayrıca otomatik olarak hiperparametre ayarı yapabilme yeteneğine sahiptir, bu da kullanıcılara modelin daha iyi bir şekilde uyarlanmasını kolaylaştırır. Bu özellikleri sayesinde,

CatBoost genellikle kategorik deęişken içeren veri setlerinde etkili bir şekilde çalışan ve rekabetçi performans gösteren bir model olarak bilinir (Rodríguez-Rodríguez vd., 2017).

## 2.10. Yapay Sinir Ağları

Son olarak, yapay sinir ağları (YSA), biyolojik sinir sisteminden ilham alarak karmaşık ilişkileri modelleyebilen bir modeldir. Bu model, derin öğrenme alanında güçlü performans gösterebilir, ancak eğitim süreleri genellikle diğer modellere göre daha uzundur (Ranjan vd., 2023).

Bu modellerin kullanılması, diyabet tahmin modelinin performansını kapsamlı bir şekilde değerlendirmek ve en etkili modeli belirlemek için çeşitli yaklaşımların bir araya getirilmesine olanak tanımaktadır.

## 3. Kullanılan Yöntem

Bu proje, bir diyabet tahmin modeli oluşturarak ve çeşitli makine öğrenimi metrikleriyle değerlendirerek, farklı modellerin performansını karşılaştırmayı amaçlamaktadır. Aşağıda, proje kapsamında yapılan işlemlerin detaylı bir açıklaması bulunmaktadır.

### 3.1. Gerekli Baęımlılıkların Projeye Eklenmesi

```
import pandas as pd
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LogisticRegression
from sklearn.metrics import accuracy_score, precision_score, recall_score, f1_score, roc_auc_score, roc_curve
from sklearn.model_selection import cross_val_score
import matplotlib.pyplot as plt
```

Şekil 1: Lojistik Regresyon Çalışması İçin Import Edilen Baęımlılıklar

İlk adım olarak projede kullanılacak tüm bağımlılıkların/kütüphanelerin kullanılabilmesi için import işlemleri yapıldı. Kullanılacak model dışındaki tüm bağımlılıklar ortak olmakla birlikte her sınıfta sadece o sınıfa ait model import edildi.

### 3.2. Veri Setinin Yüklenmesi

```
Veri

# Veri setini yükle
data = pd.read_csv('/data/diabetes.csv')
```

Şekil 2: Veri Setinin Yüklenmesi

İkinci adım olarak, projede kullanılacak olan diyabet veri seti pandas kütüphanesi kullanılarak yüklendi. Veri seti, '/data/diabetes.csv' dosyasından okunarak bir DataFrame'e dönüştürüldü.

### 3.3. Veri Setinin İncelemesi

```
# Veri setinin ilk 5 satirini incele
print(data.head())
```

[355]

...	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	\
0	6	148	72	35	0	33.6	
1	1	85	66	29	0	26.6	
2	8	183	64	0	0	23.3	
3	1	89	66	23	94	28.1	
4	0	137	40	35	168	43.1	

	DiabetesPedigreeFunction	Age	Outcome
0	0.627	50	1
1	0.351	31	0
2	0.672	32	1
3	0.167	21	0
4	2.288	33	1

Şekil 3: Veri Setinin İlk Beş Satırının İncelenmesi

```
# Veri setinin istatistikleri
data.describe().T
```

[356]

...

	count	mean	std	min	25%	50%	75%	max
Pregnancies	768.0	3.845052	3.369578	0.000	1.00000	3.00000	6.00000	17.00
Glucose	768.0	120.894531	31.972618	0.000	99.00000	117.00000	140.25000	199.00
BloodPressure	768.0	69.105469	19.355807	0.000	62.00000	72.00000	80.00000	122.00
SkinThickness	768.0	20.536458	15.952218	0.000	0.00000	23.00000	32.00000	99.00
Insulin	768.0	79.799479	115.244002	0.000	0.00000	30.50000	127.25000	846.00
BMI	768.0	31.992578	7.884160	0.000	27.30000	32.00000	36.60000	67.10
DiabetesPedigreeFunction	768.0	0.471876	0.331329	0.078	0.24375	0.37250	0.62625	2.42
Age	768.0	33.240885	11.760232	21.000	24.00000	29.00000	41.00000	81.00
Outcome	768.0	0.348958	0.476951	0.000	0.00000	0.00000	1.00000	1.00

Şekil 4: Veri Setinin İstatistiklerinin İncelenmesi

Veri setinin içeriğini anlamak adına, ilk 5 satırı ve temel istatistikleri incelendi. Bu adım, veri seti hakkında genel bir bakış sağlamak amacıyla gerçekleştirildi.

### 3.4. Bağımsız ve Bağımlı Değişkenlerin Belirlenmesi

```
# Giriş değişkenleri (bağımsız değişkenler)
X = data.drop('Outcome', axis=1)

# Çıkış değişkeni (bağımlı değişken)
y = data['Outcome']
```

Şekil 5: Bağımlı ve Bağımsız Değişkenlerin Belirlenmesi

Veri setindeki bağımsız değişkenler (X) ile bağımlı değişken (y) belirlendi. Bağımsız değişkenler, 'Outcome' sütunu hariç tüm sütunlardan oluştu.

### 3.5. Veri Setinin Eğitim ve Test Setlerine Bölünmesi

```
# Veriyi eğitim ve test setlerine bölelim
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)
```

Şekil 6: Verinin Eğitim ve Test Setlerine Bölünmesi

Makine öğrenimi modellerinin eğitilmesi ve test edilmesi için veri seti, %80 eğitim ve %20 test olacak şekilde iki ayrı sete bölündü. Bu, modelin genelleme yeteneğini değerlendirmek için önemlidir.

### 3.6. Modelin Oluşturulması ve Eğitilmesi

```
Model

# XGBoost modelini oluştur
model = XGBClassifier(n_estimators=100, random_state=42)
model.fit(X_train, y_train)
```

Şekil 7: Extreme Gradient Boosting Machines Modelinin Oluşturulması ve Eğitilmesi

Proje kapsamında on adet farklı model seçildi ve bu modeller eğitim seti üzerinde eğitildi. Şekil 7'de extreme gradient boosting machines modeli örnek olarak gösterilmiştir.

### 3.7. Tahmin Yapılması ve Performans Metriklerinin Hesaplanması

```
# Tahmin yap
y_pred = model.predict(X_test)
```



### Şekil 8: Modeli Kullanarak Tahmin Yapılması

Eğitilen model, test seti üzerinde tahminler yaparak performans metriklerini değerlendirdi. Bu metrikler arasında doğruluk (Accuracy), hassasiyet (Precision), duyarlılık (Recall), F1 puanı ve AUC (Area Under the Curve) değeri bulunmaktadır.

```
# Doğruluk (Accuracy) ölçütü
accuracy = accuracy_score(y_test, y_pred)
print('Accuracy:', accuracy)

# Cross-validation skorlarını al
cv_scores = cross_val_score(model, X_train, y_train, cv=10, scoring='accuracy')
print('Ortalama Cross-Validation Skoru:', cv_scores.mean())

# Hassasiyet (Precision) ölçütü
precision = precision_score(y_test, y_pred)
print('Precision:', precision)

# Duyarlılık (Recall) ölçütü
recall = recall_score(y_test, y_pred)
print('Recall:', recall)

# F1 puanı ölçütü
f1 = f1_score(y_test, y_pred)
print('F1 Score:', f1)
```

### Şekil 9: Performans Metriklerinin Ölçülmesi

Modelin genelleme yeteneğini daha sağlıklı bir şekilde değerlendirebilmek adına, 10 katlı cross-validation kullanıldı. Bu, modelin farklı veri alt kümeleri üzerinde nasıl performans gösterdiğini anlamak için önemlidir.

## 3.8. ROC Eğrisi ve AUC Değerinin Çizilmesi

```
# ROC eğrisi ve AUC değeri
y_probs = model.predict_proba(X_test)[: , 1]
fpr, tpr, thresholds = roc_curve(y_test, y_probs)
roc_auc = roc_auc_score(y_test, y_probs)
```

### Şekil 10: ROC Eğrisinin ve AUC Değerinin Çizilmesi

Modelin sınıflandırma performansını daha iyi anlamak ve karşılaştırmak amacıyla ROC (Receiver Operating Characteristic) eğrisi çizildi. AUC değeri, bu eğri altındaki alanı ifade ederek modelin sınıflandırma performansını ölçer.

### 3.9. Grafikselle Gösterimler ile Sonuçların İfade Edilmesi

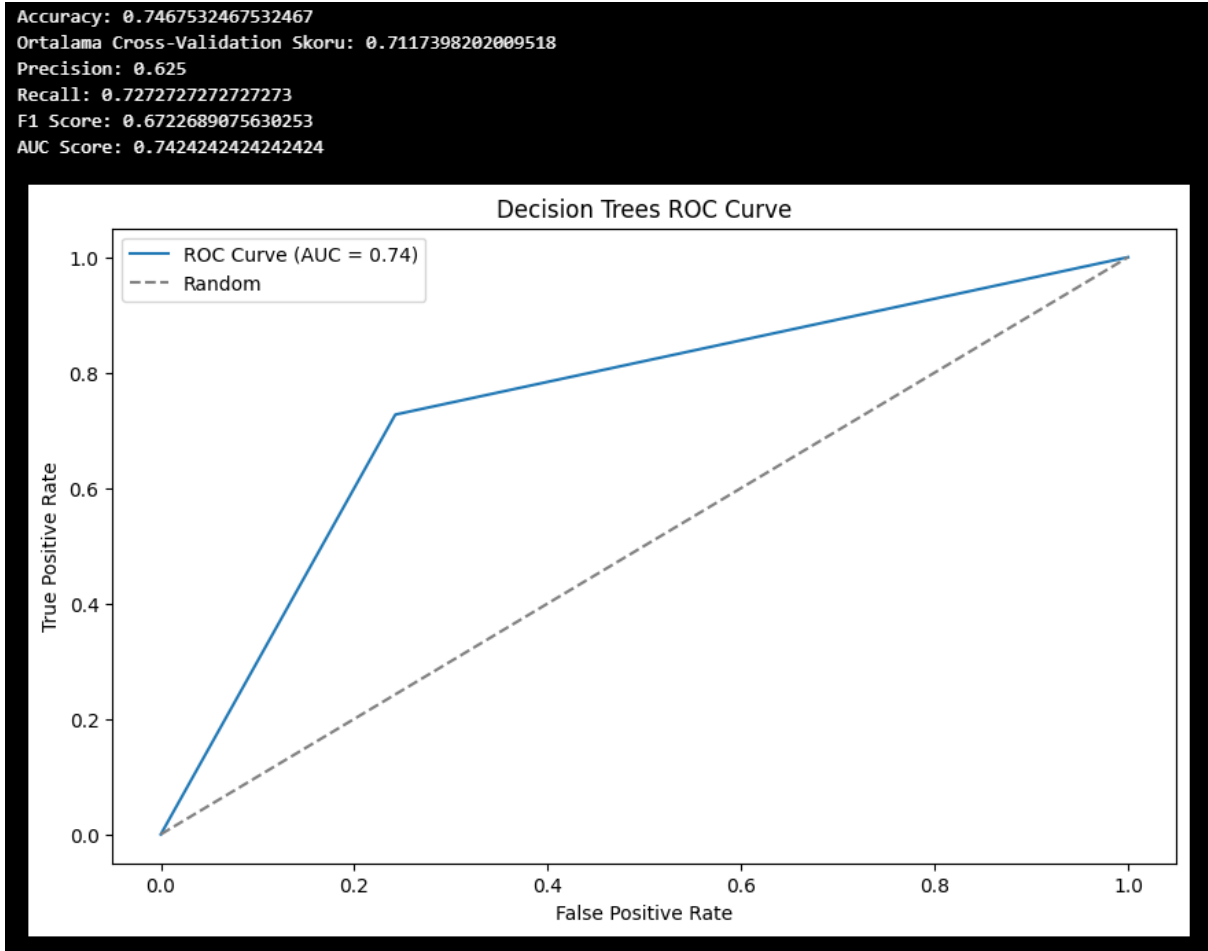
```
# ROC eğrisini çiz
plt.figure(figsize=(10, 6))
plt.plot(fpr, tpr, label='ROC Curve (AUC = {:.2f})'.format(roc_auc))
plt.plot([0, 1], [0, 1], linestyle='--', color='gray', label='Random')
plt.xlabel('False Positive Rate')
plt.ylabel('True Positive Rate')
plt.title('Logistic Regression ROC Curve')
plt.legend()
plt.show()
```

Şekil 11: ROC Eğrisinin Grafikselle Olarak Gösterilmesi

ROC eğrisi ve diğer performans metrikleri, grafikler aracılığıyla görsel olarak ifade edilerek projenin sonuçları daha anlaşılır hale getirildi. Bu görsel ifadeler, modelin performansını açıklamada ve karşılaştırmada yardımcı olması sebebiyle önemlidir.

## 4. Elde Edilen Sonuçların Değerlendirilmesi

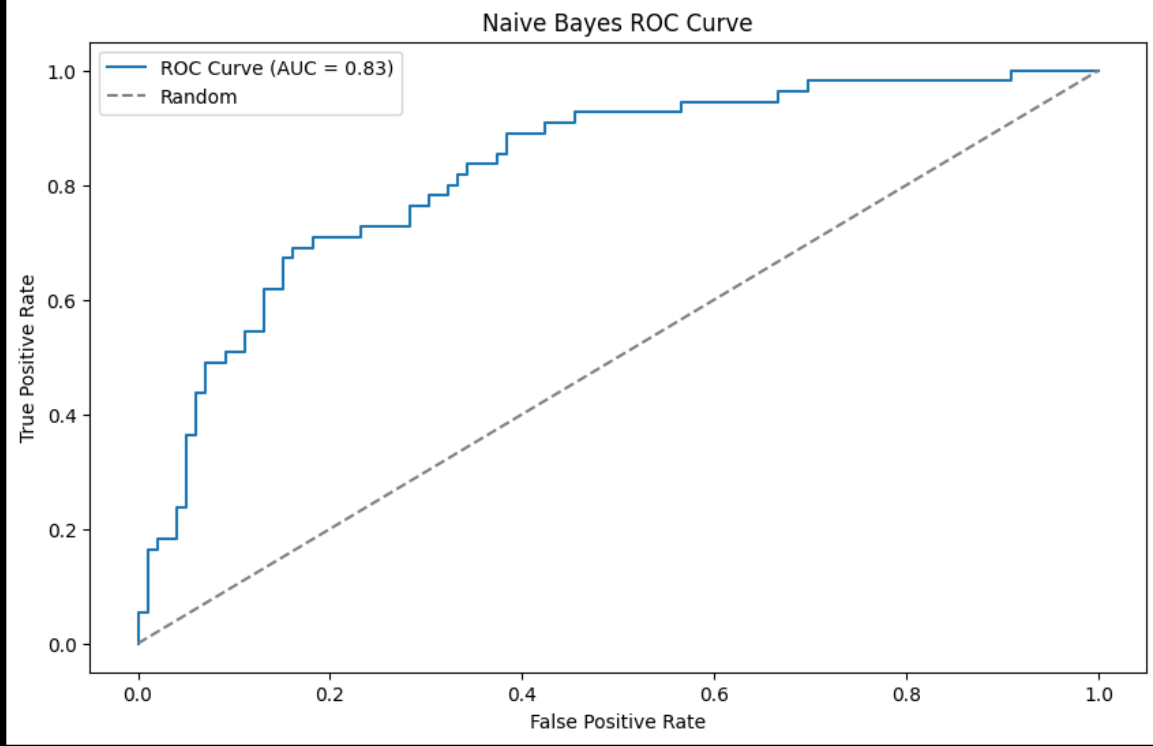
### 4.1. Bulgular



Şekil 12: Karar Ağaçları Modelinin Performansı

Bu çalışmada kullanılan karar ağaçları modeli, diyabet tahmininde iyi bir performans sergilemiştir. Modelin doğruluk oranı %74.68 olarak ölçüldü. Cross-validation sonuçlarına göre, modelin genelleme yeteneği de göz önüne alındığında, ortalama cross-validation skoru %71.17 olarak belirlendi. Hassasiyet değeri %62.50, duyarlılık değeri ise %72.73 olarak hesaplandı. F1 skoru ise %67.23 olarak ölçüldü. Ayrıca, ROC eğrisi altındaki alan (AUC) değeri %74.24 olarak belirlendi.

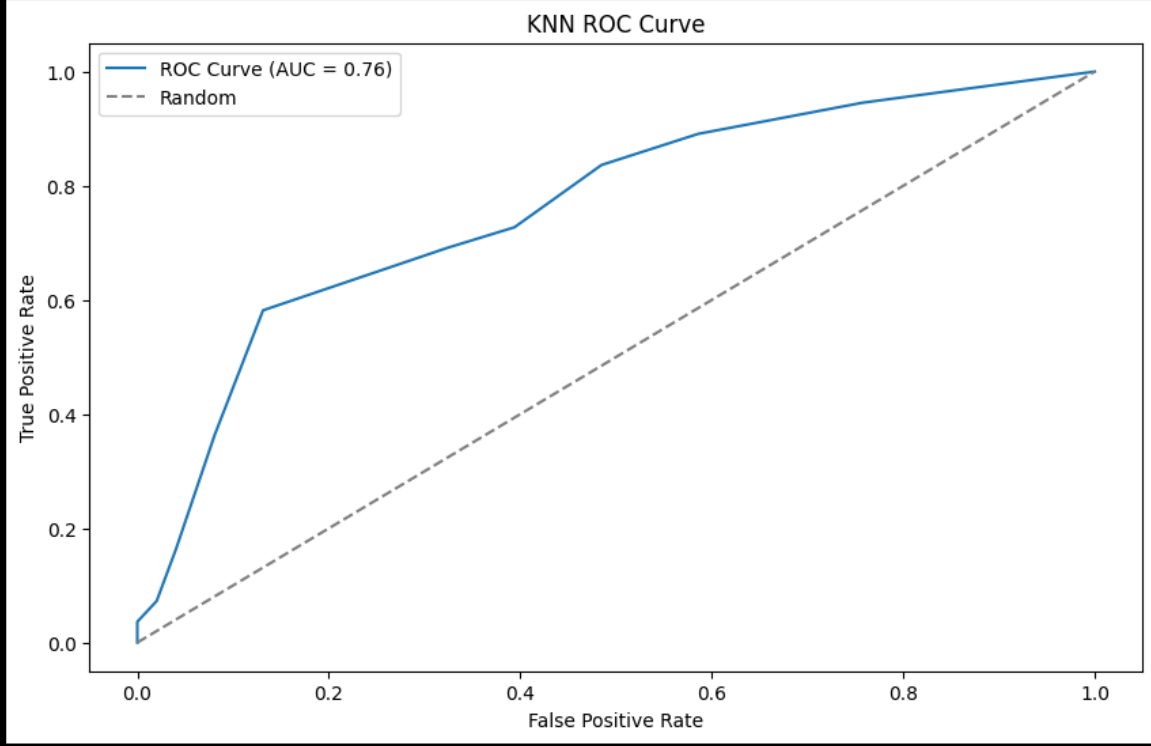
Accuracy: 0.7662337662337663  
Ortalama Cross-Validation Skoru: 0.7475674246430459  
Precision: 0.6610169491525424  
Recall: 0.7090909090909091  
F1 Score: 0.6842105263157895  
AUC Score: 0.8253443526170798



Şekil 13: Naive Bayes Modelinin Performansı

Naive Bayes modeli, diyabet tahmininde karar ağaçlarına benzer bir performans sergiledi. Doğruluk oranı %76.62 olarak ölçüldü. Cross-validation sonuçlarına göre, modelin genelleme yeteneği ortalama cross-validation skoru %74.76 oldu. Hassasiyet değeri %66.10, duyarlılık değeri ise %70.91 olarak hesaplandı. F1 skoru %68.42, AUC değeri ise %82.53 olarak belirlendi.

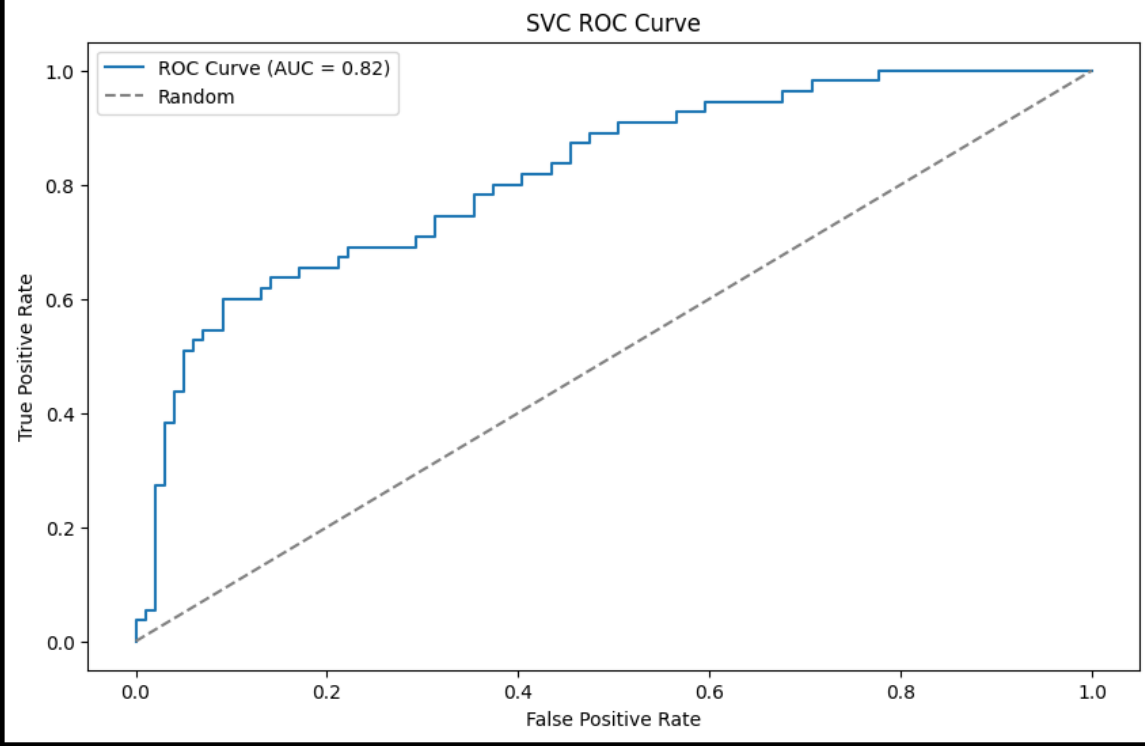
Accuracy: 0.7662337662337663  
Ortalama Cross-Validation Skoru: 0.7378635642517187  
Precision: 0.7111111111111111  
Recall: 0.5818181818181818  
F1 Score: 0.64  
AUC Score: 0.7620752984389347



Şekil 14: K-En Yakın Komşu Modelinin Performansı

KNN modeli, diyabet tahmininde yüksek bir doğruluk oranı olan %76.62'lik bir sonuç elde etti. Cross-validation skorlarına göre genelleme yeteneği, ortalama cross-validation skoru %73.79 olarak ölçüldü. Hassasiyet değeri %71.11, duyarlılık değeri ise %58.18 olarak hesaplandı. F1 skoru %64.00, AUC değeri ise %76.21 olarak belirlendi.

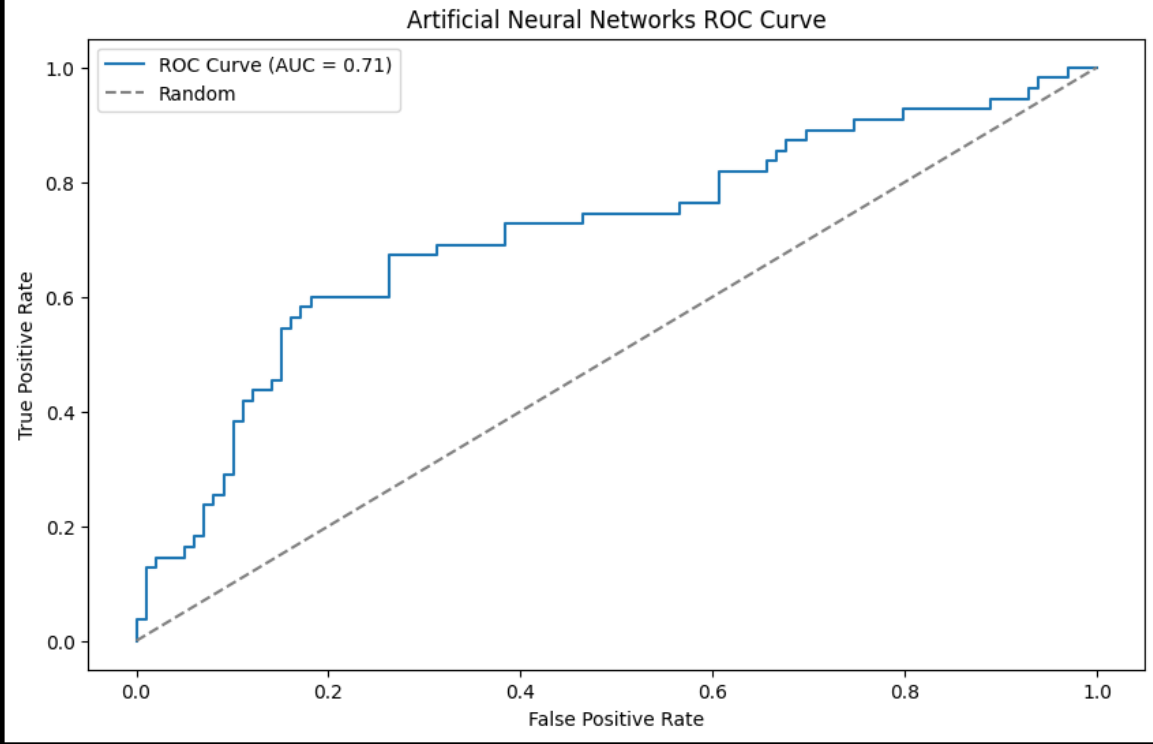
Accuracy: 0.7532467532467533  
Ortalama Cross-Validation Skoru: 0.7703067160232682  
Precision: 0.6545454545454545  
Recall: 0.6545454545454545  
F1 Score: 0.6545454545454545  
AUC Score: 0.8165289256198347



Şekil 15: Destek Vektör Makineleri Modelinin Performansı

SVM modeli, diyabet tahmininde %75.32'lik bir doğruluk oranına sahiptir. Cross-validation sonuçlarına göre, modelin genelleme yeteneği ortalama cross-validation skoru %77.03'tür. Hassasiyet değeri %65.45, duyarlılık değeri ise %65.45 olarak ölçüldü. F1 skoru %65.45, AUC değeri ise %81.65 olarak belirlendi.

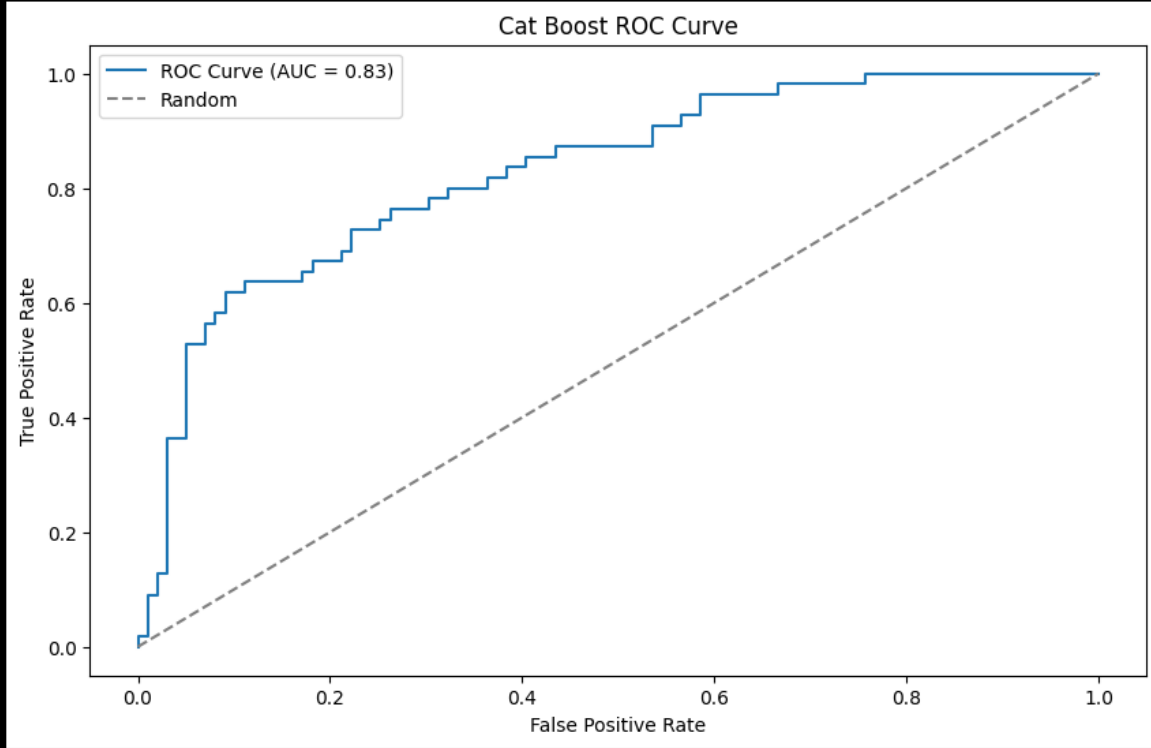
Accuracy: 0.6818181818181818  
Ortalama Cross-Validation Skoru: 0.6842675832892648  
Precision: 0.5441176470588235  
Recall: 0.6727272727272727  
F1 Score: 0.6016260162601627  
AUC Score: 0.7114784205693296



Şekil 16: Yapay Sinir Ağları Modelinin Performansı

Yapay sinir ağları modeli, diyabet tahmininde %68.18'lik bir doğruluk oranı elde etti. Cross-validation sonuçlarına göre, modelin genelleme yeteneği ortalama cross-validation skoru %68.43 olarak belirlendi. Hassasiyet değeri %54.41, duyarlılık değeri ise %67.27 olarak ölçüldü. F1 skoru %60.16, AUC değeri ise %71.15 olarak hesaplandı.

Accuracy: 0.7597402597402597  
Ortalama Cross-Validation Skoru: 0.7685880486515071  
Precision: 0.6666666666666666  
Recall: 0.6545454545454545  
F1 Score: 0.6605504587155963  
AUC Score: 0.8288337924701561

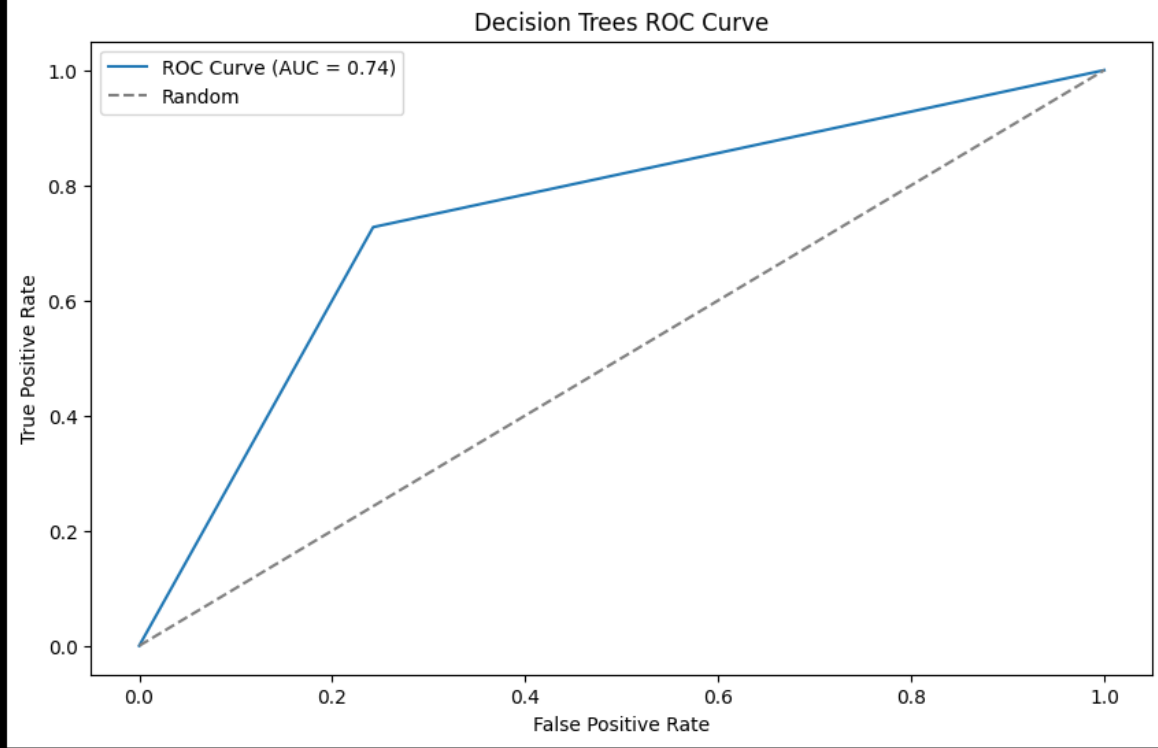


Şekil 17: CatBoost Modelinin Performansı

CatBoost modeli, diyabet tahmininde %75.97'lik bir doğruluk oranına sahiptir. Cross-validation sonuçlarına göre, modelin genelleme yeteneği ortalama cross-validation skoru %76.86'dır. Hassasiyet değeri %66.67, duyarlılık değeri ise %65.45 olarak ölçüldü. F1 skoru %66.06, AUC değeri ise %82.88 olarak belirlendi.



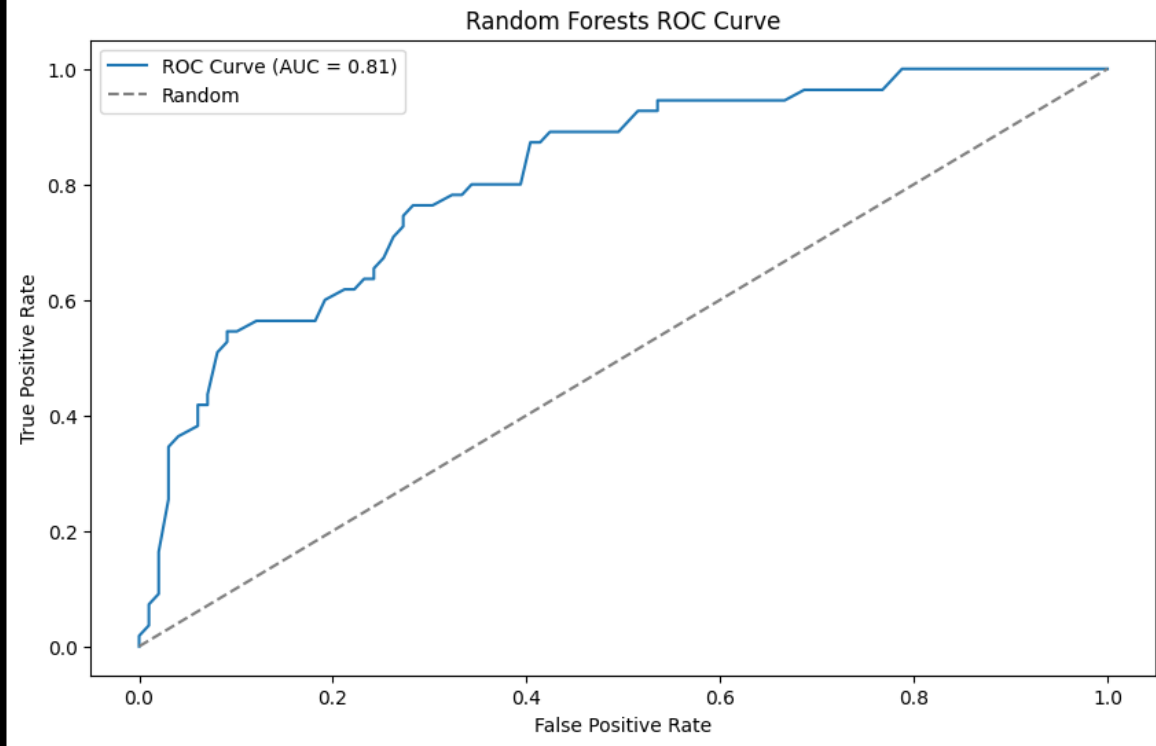
Accuracy: 0.7467532467532467  
Ortalama Cross-Validation Skoru: 0.7117398202009518  
Precision: 0.625  
Recall: 0.7272727272727273  
F1 Score: 0.6722689075630253  
AUC Score: 0.7424242424242424



Şekil 18: Karar Ağaçları Modelinin Performansı

Karar ağaçları modeli, diyabet tahmininde %74.68'lik bir doğruluk oranı elde etti. Cross-validation sonuçlarına göre, modelin genelleme yeteneği ortalama cross-validation skoru %71.17 olarak belirlendi. Hassasiyet değeri %62.50, duyarlılık değeri ise %72.73 olarak ölçüldü. F1 skoru %67.23, AUC değeri ise %74.24 olarak hesaplandı.

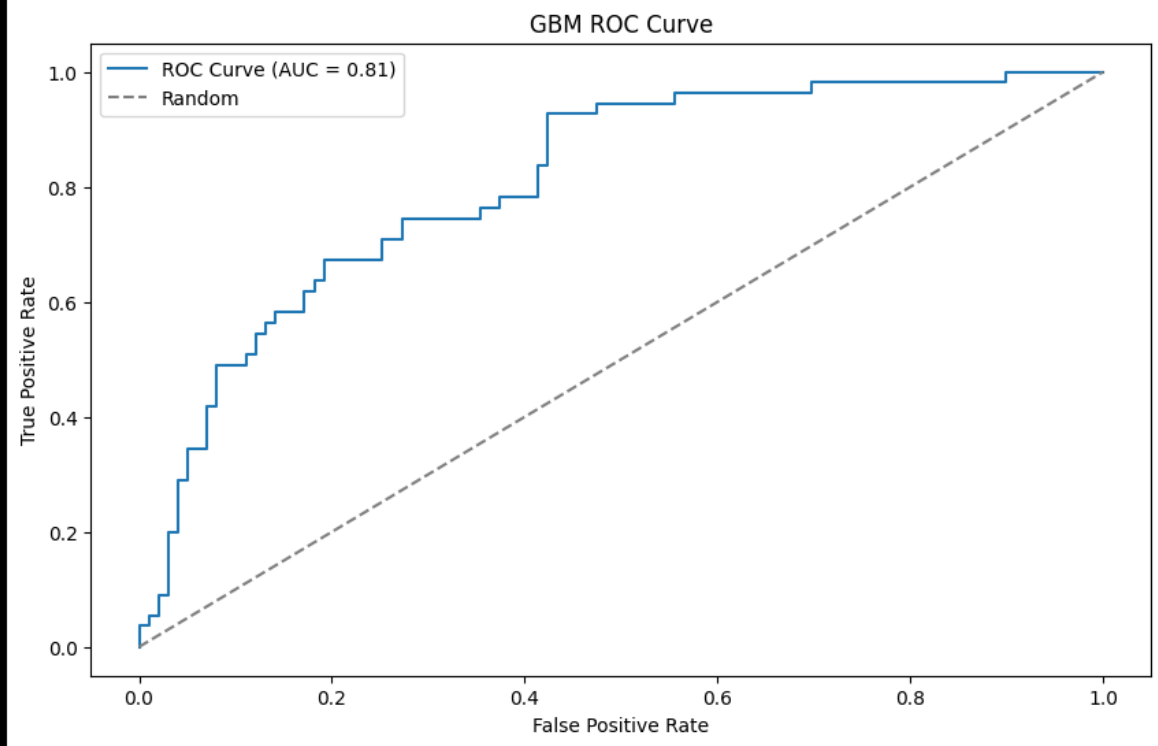
Accuracy: 0.7207792207792207  
Ortalama Cross-Validation Skoru: 0.7637493389740877  
Precision: 0.6071428571428571  
Recall: 0.6181818181818182  
F1 Score: 0.6126126126126126  
AUC Score: 0.8120293847566575



Şekil 19: Random Forests Modelinin Performansı

Random Forests modeli, diyabet tahmininde %72.08'lik bir doğruluk oranına sahiptir. Cross-validation sonuçlarına göre, modelin genelleme yeteneği ortalama cross-validation skoru %76.37'dir. Hassasiyet değeri %60.71, duyarlılık değeri ise %61.82 olarak ölçüldü. F1 skoru %61.26, AUC değeri ise %81.20 olarak belirlendi.

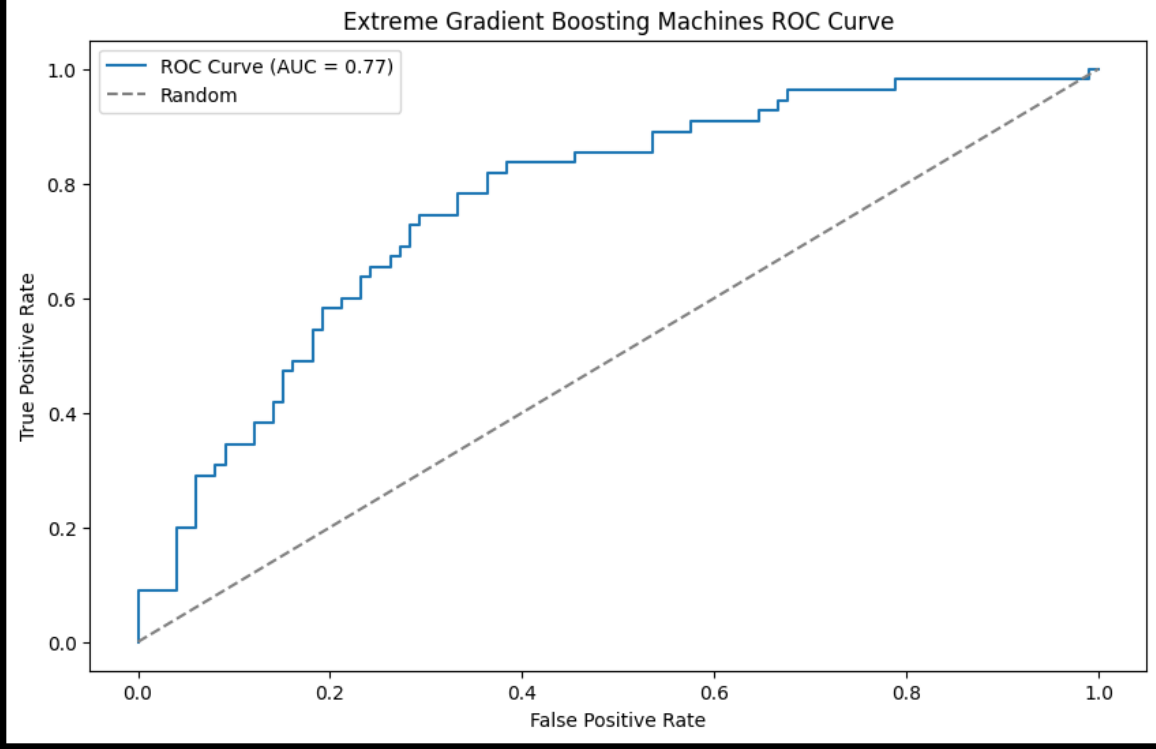
Accuracy: 0.7467532467532467  
Ortalama Cross-Validation Skoru: 0.7734796404019038  
Precision: 0.6379310344827587  
Recall: 0.6727272727272727  
F1 Score: 0.6548672566371682  
AUC Score: 0.8091827364554638



Şekil 20: Gradient Boosting Machines Modelinin Performansı

Gradient Boosting Machines modeli, diyabet tahmininde %74.68'lik bir doğruluk oranı elde etti. Cross-validation sonuçlarına göre, modelin genelleme yeteneği ortalama cross-validation skoru %77.35'tir. Hassasiyet değeri %63.79, duyarlılık değeri ise %67.27 olarak ölçüldü. F1 skoru %65.49, AUC değeri ise %80.92 olarak hesaplandı.

Accuracy: 0.7077922077922078  
Ortalama Cross-Validation Skoru: 0.7686673717609731  
Precision: 0.5806451612903226  
Recall: 0.6545454545454545  
F1 Score: 0.6153846153846153  
AUC Score: 0.766574839302112

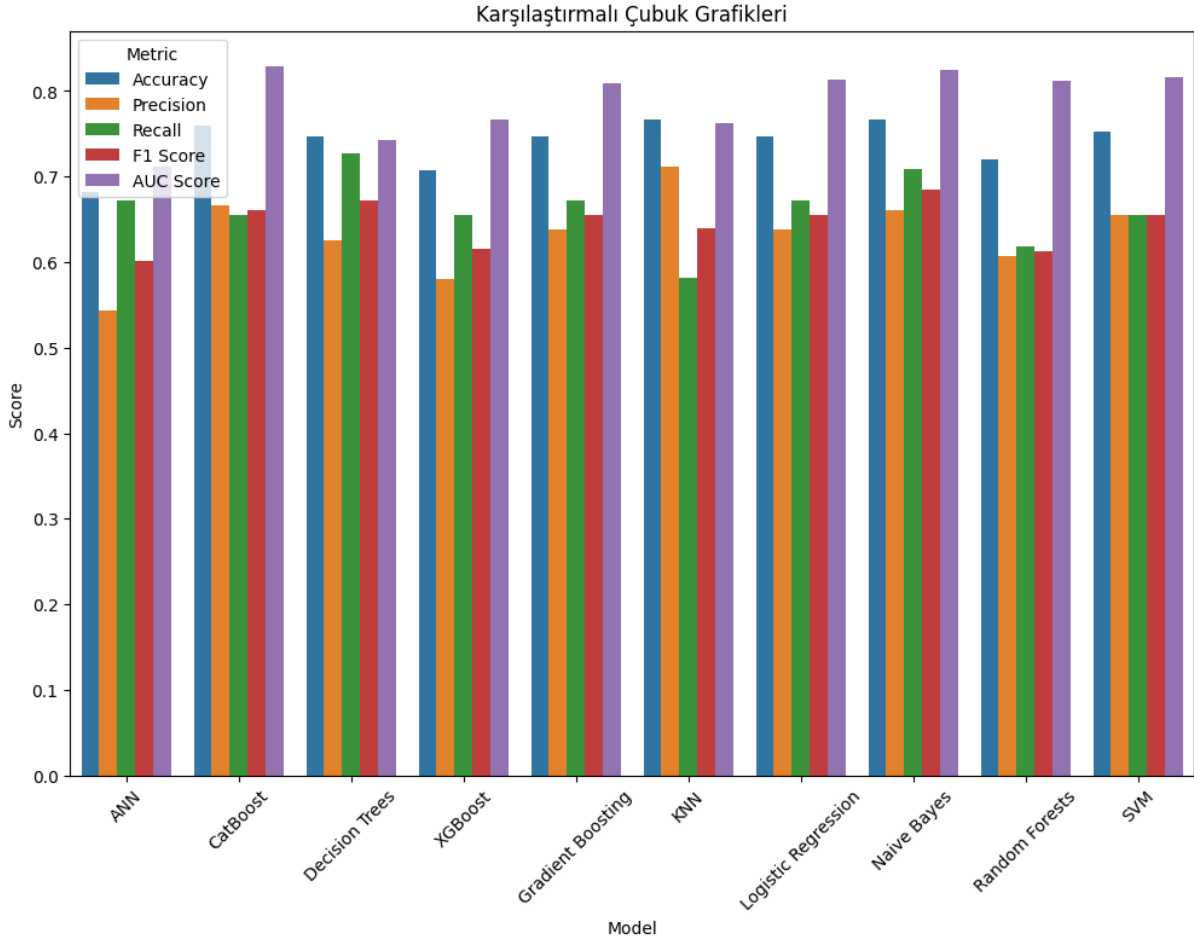


Şekil 21: Extreme Gradient Boosting Machines (XGBoost) Modelinin Performansı

XGBoost modeli, diyabet tahmininde %70.78'lik bir doğruluk oranı elde etti. Cross-validation sonuçlarına göre, modelin genelleme yeteneği ortalama cross-validation skoru %76.87'dir. Hassasiyet değeri %58.06, duyarlılık değeri ise %65.45 olarak ölçüldü. F1 skoru %61.54, AUC değeri ise %76.66 olarak belirlendi.

Bu sonuçlar, farklı makine öğrenimi modellerinin performansını değerlendirmek için çeşitli ölçütleri kullanarak elde edildi. Her bir modelin avantajları ve zayıf yönleri, doğruluk, hassasiyet, duyarlılık, F1 skoru ve AUC değeri gibi metrikler üzerinden karşılaştırılarak belirlendi. Bir sonraki bölümde bu sonuçların bir analizi yapılacaktır.

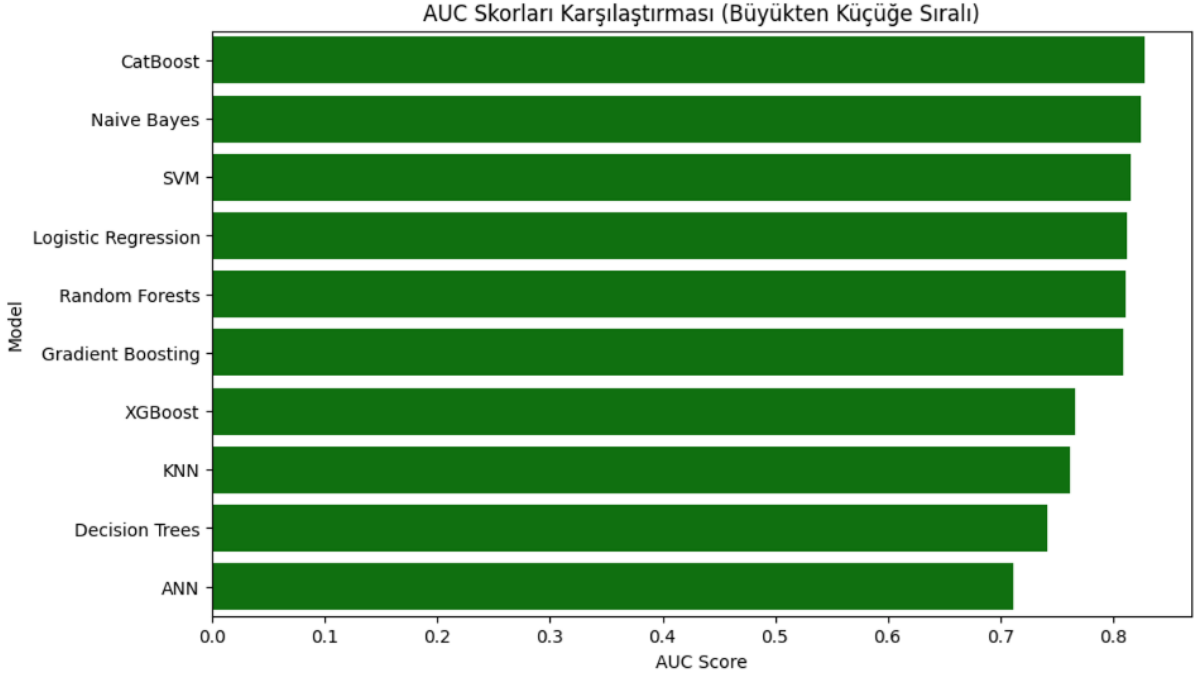
## 4.2. Bulguların Yorumlanması ve Sonuç



Şekil 22: Tüm Performans Metriklerinin Bir Arada Gösterimi

Bu çalışma, diyabet tahmini için çeşitli makine öğrenimi modellerini değerlendirmeyi amaçlamıştır. İncelenen modeller arasında genel bir başarı trendi gözlemlenmiştir.

En yüksek AUC skoruna sahip model, CatBoost olmuştur (%82.88). CatBoost'un yüksek AUC skoru, modelin sınıflandırma performansının genel olarak üst düzeyde olduğunu göstermektedir.



Şekil 23: Tüm Modellerin AUC Skorları (Sıralı)

Kullanılan makine öğrenmesi modelleri arasında extreme gradient boost machines, k-en yakın komşu ve karar ağaçları modelleri, diğer modellere göre daha zayıf bir performans göstermiştir.

Buna karşılık, yapay sinir ağları modeli (%68.18 doğruluk, %71.15 AUC) en düşük doğruluk oranına ve AUC skoruna sahiptir. Yapay sinir ağları, genelleme yeteneği konusunda diğer modellere göre biraz zayıf kalmıştır. Daha karmaşık yapılar ve hiperparametre ayarlaması ile performansının artırılması mümkün olabilir.

Genel olarak, modellerin performansı değerlendirilirken doğruluk, hassasiyet, duyarlılık, F1 skoru ve AUC gibi çeşitli ölçütler dikkate alınmalıdır. Model seçimi, spesifik bir problemin gereksinimleri ve veri setinin özellikleri göz önüne alınarak yapılmalıdır. CatBoost, genel performansı ile öne çıksa da, her modelin avantajları ve zayıf yönleri detaylı bir analizle değerlendirilmelidir.

Sonuç olarak, bu çalışma, farklı makine öğrenimi modellerinin diyabet tahminindeki başarılarını karşılaştırarak, hangi modellerin daha etkili olduğunu anlamak için bir temel sağlamaktadır. İlerleyen araştırmalarda, daha büyük veri setleri ve daha fazla hiperparametre ayarlaması ile modellerin performansının daha da artırılması hedeflenebilir.

## KAYNAKLAR

Baher, H. L., Lemaire, V., & Trinquart, R. (2020). On the intrinsic robustness to noise of some leading classifiers and symmetric loss function--an empirical evaluation. arXiv preprint arXiv:2010.13570.

Bonfietti, A., Lombardi, M., & Milano, M. (2015). Embedding decision trees and random forests in constraint programming. *Integration of AI and OR Techniques in Constraint Programming*, 74–90. [https://doi.org/10.1007/978-3-319-18008-3\\_6](https://doi.org/10.1007/978-3-319-18008-3_6)

Dorogush, A. V., Ershov, V., & Gulin, A. (2018). CatBoost: gradient boosting with categorical features support. arXiv preprint arXiv:1810.11363.

Ganesh, V., Kolluri, J., & Kumar, K. V. (2021). Diabetes prediction using logistic regression and feature normalization. *2021 International Conference on Innovative Computing, Intelligent Communication and Smart Electrical Systems (ICSSES)*. <https://doi.org/10.1109/icses52305.2021.9633773>

Gupta, S. C., & Goel, N. (2020). Enhancement of performance of K-nearest neighbors classifiers for the prediction of diabetes using feature selection method. *2020 IEEE 5th International Conference on Computing Communication and Automation (ICCCA)*, 681–686. <https://doi.org/10.1109/iccca49541.2020.9250887>

Hasanah, W., & Munggaran, L. C. (2021). Comparison of naïve Bayes and Random Forest Methods for diabetes prediction. *International Journal of Computer Applications*, 174(26), 13–18. <https://doi.org/10.5120/ijca2021921184>

Jain, V. (2022). Diabetes prediction using support vector machine, naive Bayes and Random Forest Machine Learning Models. 2022 6th International Conference on Electronics, Communication and Aerospace Technology, 837–841. <https://doi.org/10.1109/iceca55336.2022.10009241>

Joshi, T.N., & Chawan, P. (2018). LOGISTIC REGRESSION AND SVM BASED DIABETES PREDICTION SYSTEM.

Kapoor, S., & Perrone, V. (2021). A Simple and Fast Baseline for Tuning Large XGBoost Models. arXiv preprint arXiv:2111.06924.

Liu, W., Zhang, L., Xin, Z., Zhang, H., You, L., Bai, L., Zhou, J., & Ying, B. (2022). A promising preoperative prediction model for microvascular invasion in hepatocellular carcinoma based on an extreme gradient boosting algorithm. *Frontiers in Oncology*, 12. <https://doi.org/10.3389/fonc.2022.852736>

Marzi, C., d'Ambrosio, A., Diciotti, S., Bisecco, A., Altieri, M., Filippi, M., Rocca, M. A., Storelli, L., Pantano, P., Tommasin, S., Cortese, R., De Stefano, N., Tedeschi, G., & Gallo, A. (2023). Prediction of the information processing speed performance in multiple sclerosis using a machine learning approach in a large Multicenter Magnetic Resonance Imaging Data Set. *Human Brain Mapping*, 44(1), 186–202. <https://doi.org/10.1002/hbm.26106>



Okikiola, F. M., Adewale, O. S., & Obe, O. O. (2023). A diabetes prediction classifier model using naive Bayes algorithm. *FUDMA JOURNAL OF SCIENCES*, 7(1), 253–260. <https://doi.org/10.33003/fjs-2023-0701-1301>

Prakash, A., Anand, R., Abinayaa, S. S., & Kalyan Chakravarthy, N. S. (2021). Normalized naïve bayes model to predict type –2 diabetes mellitus. 2021 Emerging Trends in Industry 4.0 (ETI 4.0). <https://doi.org/10.1109/eti4.051663.2021.9619332>

Ranjan, R., Gupta, S., Sharma, S., Alatba, S. R., Sreeram, A., & Fernandez, S. (2023). Artificial neural network based categorization of diabetes mellitus and its comparison with K-nearest neighbors algorithm. 2023 3rd International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE), 1099–1101. <https://doi.org/10.1109/icacite57410.2023.10182961>

Rodríguez-Rodríguez, I., Zamora, M. Á., & Rodríguez, J.-V. (2017). On predicting glycaemia in type 1 diabetes mellitus patients by using support vector machines. Proceedings of the 1st International Conference on Internet of Things and Machine Learning. <https://doi.org/10.1145/3109761.3158404>

Sinaga, L. M., Sawaluddin, & Suwilo, S. (2020). Analysis of classification and naïve Bayes algorithm K-nearest neighbor in data mining. *IOP Conference Series: Materials Science and Engineering*, 725(1), 012106. <https://doi.org/10.1088/1757-899x/725/1/012106>

Solichin, M., Nurahmi, L., & Putro, B. J. (2019). Power prediction of a 4-CRU parallel mechanism based on extra gradient boosting regressor. *AIP Conference Proceedings*. <https://doi.org/10.1063/1.5138345>

Zhu, C., Idemudia, C. U., & Feng, W. (2019). Improved logistic regression model for diabetes prediction by integrating PCA and K-means techniques. *Informatics in Medicine Unlocked*, 17, 100179. <https://doi.org/10.1016/j.imu.2019.100179>