

**T.C.**

**KONYA TEKNİK ÜNİVERSİTESİ**

**Veri Madenciliği Dersi**

**Proje Adı:**

**Diyabet Hastalığının Verisel Ön Teşhisi/Tanısı**

**Öğrenci Adı:**

**Elife Beyzanur YÜKSEL**

**Öğrenci No:**

**171213069**

**2024**

**KONYA**

**Özet**

Bu çalışma, hastaların diyabet durumunu öngörme amacıyla makine öğrenimi modellerini değerlendirmiştir. Veri seti analizi sonucunda, glikoz değeri, vücut kitle indeksi (BMI) ve yaş gibi faktörlerin diyabet üzerinde belirleyici olduğu tespit edilmiştir. Bu doğrultuda, lojistik regresyon, destek vektör makineleri (SVM), Naive Bayes, karar ağaçları ve yapay sinir ağları gibi çeşitli modeller eğitilmiş ve değerlendirilmiştir.

Deneyler sonucunda, Naive Bayes modelinin en yüksek doğruluk skoruna sahip olduğu belirlenmiştir. Bu model, diyabet riskini değerlendirmek için güvenilir bir araç olarak öne çıkmaktadır. Ayrıca, eğitilmiş modelin "diabetes.sav" dosyasına kaydedilerek gelecekteki çalışmalarda veya uygulamalarda kullanılması için saklandığı tespit edilmiştir.

Bu çalışma, makine öğrenimi tekniklerinin diyabet teşhisi koymada önemli bir potansiyele sahip olduğunu göstermektedir. Erken teşhis, hastaların sağlık durumunu izlemek ve uygun tedaviyi başlatmak için kritik öneme sahiptir. Bu nedenle, diyabetin erken teşhisi için bu tür modellerin kullanımı sağlık alanında önemli bir adım olabilir.

İçindekiler:

1. Proje Amacı: Hastaların Diyabet Durumunu Öngörme
2. Veri Seti

2.1. Veri Seti Özellikleri

2.2. Veri Setinin Özelliklere Göre Isı Dağılım Haritas**ı**

2.3. Veri Seti Çıkarımları

2.3.1. Diyabete Etki Eden En Önemli Faktörler

2.3.1.1. Glikoz Değeri

2.3.1.2. Vücut Kitle Endeksi (BMI)

2.3.1.3. Yaş

2.3.2. Diyabete Etkisi En Az Olan Faktörler

2.3.2.1. Diyabetik Soy Ağacı (DiabetesPedigreeFunction):

2.3.2.2. Tansiyon (BloodPressure)

2.3.2.3. Hamilelik Sayısı (Pregnancies):

1. Grafikler

3.1. Kabarcık Grafiği

3.2. Çubuk Grafiği

1. Model

4.1. Modeller

4.1.1. Lojistik Regresyon

4.1.2. Destek Vektör Makineleri (SVM)

4.1.3. Naive Bayes

4.1.4. Karar Ağacı

4.1.5. Yapay Sinir Ağı

4.2. Modellerin Eğitimi ve Test Gruplarının Oluşturulması

1. Eğitimler

5.1. Modellerin Eğitime Göre Sonuçlarının Değerlendirilmesi

5.1.1. Eğitim Verisinin Bölünmeden Kullanıldığı Durum

5.1.2. Eğitim Verisinin Bölünerek Kullanıldığı Durum

5.2. Eğitim Verilerinin Kaydedilmesi

5.3. Eğitilmiş Makinenin Test Edilmesi

1. Sonuç
2. Kaynakça

1. **Proje Amacı: Hastaların Diyabet Durumunu Öngörme**

Bu proje, hastaların belirli özelliklerini kullanarak (yaş, vücut kitle indeksi, kan şekeri seviyeleri vb.), hastaların diyabet olup olmadığını belirlemeyi amaçlamaktadır. Diyabet, dünya genelinde yaygın bir sağlık sorunudur ve erken teşhis edilmesi ve tedavi edilmesi önemlidir. Bu nedenle, bu projenin amacı, hastaların diyabet riskini değerlendirmek için veri analitiği ve makine öğrenimi tekniklerini kullanmaktır.

Proje, farklı makine öğrenimi algoritmalarını (Lojistik Regresyon, Destek Vektör Makineleri, Naive Bayes, Karar Ağacı ve Yapay Sinir Ağları gibi) kullanarak hastaların diyabet durumunu belirlemek için bir model geliştirmeyi hedeflemektedir. Bu algoritmalar, hastaların sağlık verilerine dayalı olarak diyabet riskini değerlendirmek için farklı yaklaşımlar sunmaktadır.

Sonuç olarak, bu proje, diyabet hastalığının erken teşhisine katkıda bulunmayı amaçlamakta ve hastaların sağlık durumunu değerlendirmek için veri analitiği ve makine öğrenimi tekniklerini kullanmaktadır.

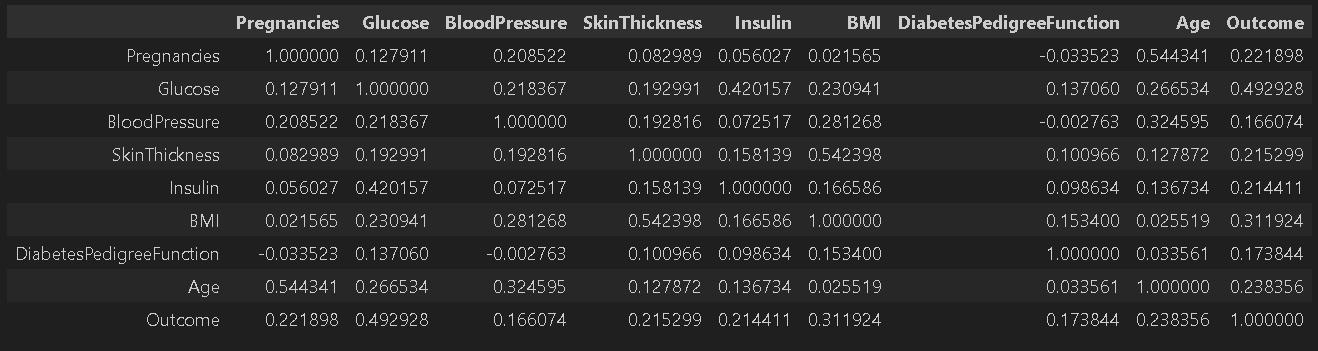
1. **Veri Seti**

**2.1. Veri Seti Özellikleri**

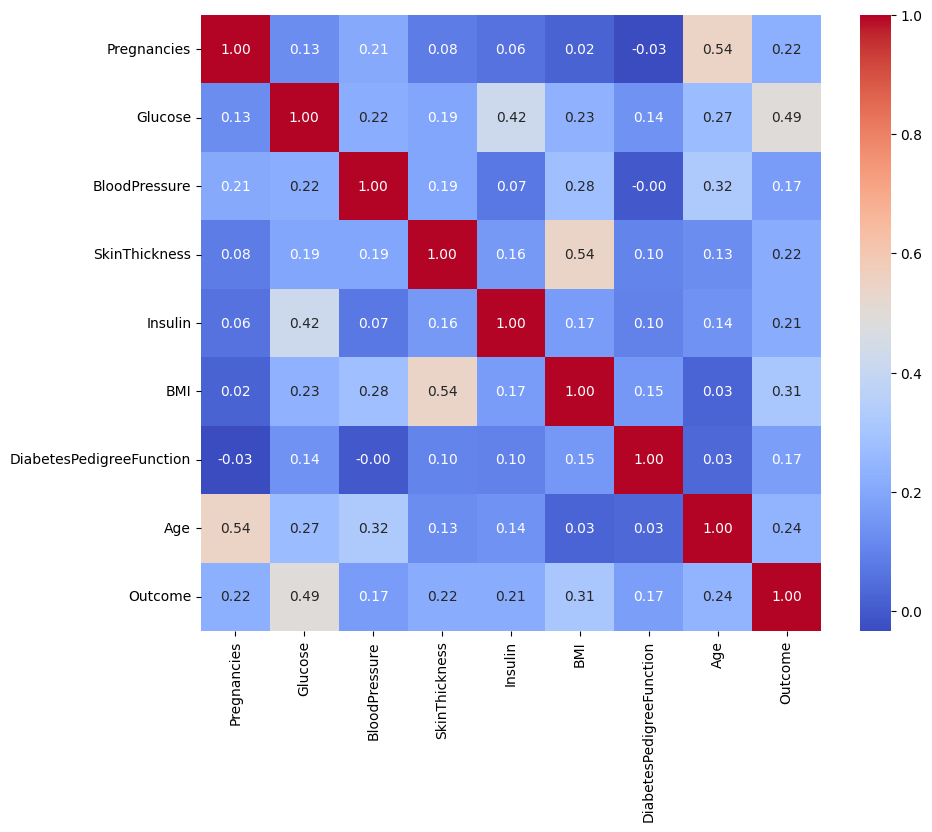
Veri setimiz, hastaların diyabet durumunu belirlemek için kullanılan 9 farklı özelliği içermektedir. Bu özellikler hamilelik sayısı, glikoz seviyeleri, tansiyon, cilt kalınlığı, insülin seviyeleri, vücut kitle indeksi (BMI), diyabet soyağacı fonksiyonu, yaş ve sonuç (diyabet olma durumu) özelliklerini kapsamaktadır. Veri seti, toplamda 768 örnekten oluşmaktadır.

**2.2. Veri Setinin Özelliklere Göre Isı Dağılım Haritası**

Veri setinin özellikleri arasındaki ilişkileri anlamak için korelasyon analizi yapılmış ve bu analizin sonuçları bir ısı haritası şeklinde görselleştirilmiştir. Aşağıda, özellikler arasındaki korelasyon katsayıları verilmiştir:



Bu korelasyon analizi yapılmış olan tablonun ısı dağılım grafiği oluşturulmuştur.



**2.3. Veri Seti Çıkarımları**

**2.3.1. Diyabete Etki Eden En Önemli Faktörler**

Isı haritasına göre, diyabete etkileyen en önemli üç faktör şunlardır:

**2.3.1.1. Glikoz Değeri:** Glikoz seviyesi, diyabet riski üzerinde en güçlü etkiye sahip özelliktir. Glikoz seviyesi ile diyabet arasındaki korelasyon katsayısı 0.49'dur, yani bu özellik diyabet olma olasılığını oldukça etkilemektedir.

**2.3.1.2. Vücut Kitle İndeksi (BMI):** Vücut kitle indeksi (BMI), diyabet riski üzerinde ikinci en güçlü etkiye sahip özelliktir. BMI ile diyabet arasındaki korelasyon katsayısı 0.31'dir. Bu, kişinin vücut ağırlığına göre belirlenen bir ölçü olan BMI'nin diyabet olma olasılığıyla ilişkili olduğunu göstermektedir.

**2.3.1.3. Yaş:** Yaş, diyabet riski üzerinde önemli bir faktördür. Yaş ile diyabet arasındaki korelasyon katsayısı 0.24'tür. Bu, yaşın artmasıyla diyabet olma olasılığının arttığını göstermektedir

Bu bulgular, diyabet riskini belirlemede önemli olan faktörleri anlamamıza yardımcı olmaktadır. Glikoz seviyesi, vücut kitle indeksi ve yaş, diyabet riskini değerlendirirken dikkate alınması gereken anahtar faktörlerdir.

**2.3.2. Diyabete Etkisi En Az Olan Faktörler:**Isı haritasına göre, diyabete etkileyen en az etkili üç faktör şunlardır:

**2.3.2.1. Diyabetik Soy Ağacı (DiabetesPedigreeFunction):** Diyabetik soy ağacı, diyabet riski üzerinde en az etkiye sahip faktörlerden biridir. Korelasyon katsayısı 0.17 olan bu özellik, genetik geçişin diyabet olasılığı üzerindeki göreceli olarak düşük etkisini yansıtmaktadır.

**2.3.2.2. Tansiyon (BloodPressure):** Tansiyon seviyeleri, diyabet riski üzerinde de sınırlı bir etkiye sahiptir. Korelasyon katsayısı 0.17 olan bu özellik, tansiyonun diyabet olasılığına doğrudan etkisinin düşük olduğunu göstermektedir.

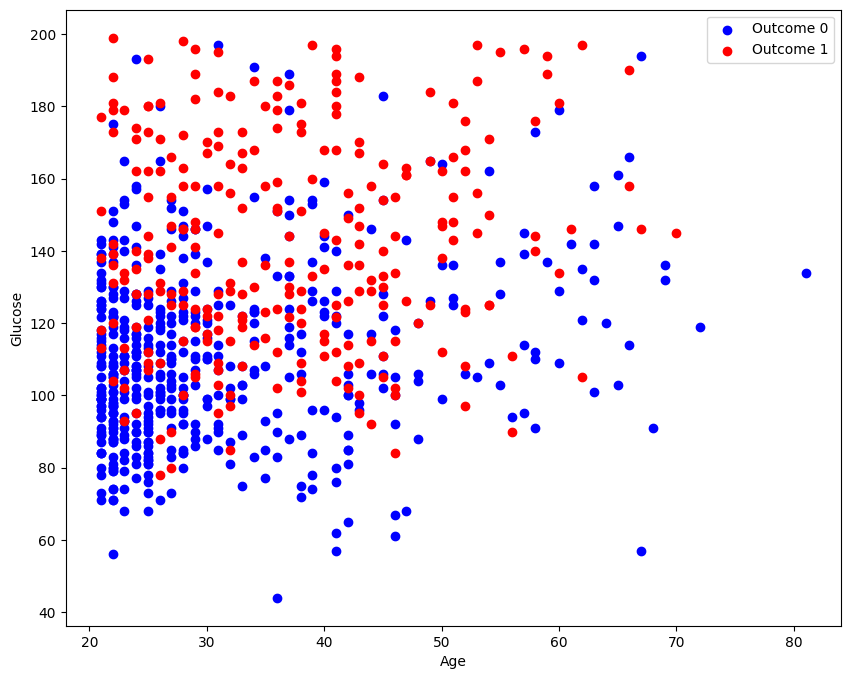
**2.3.2.3. Hamilelik Sayısı (Pregnancies):** Hamilelik sayısı, diyabet riski üzerindeki etkisi en düşük olan faktörlerden biridir. Korelasyon katsayısı 0.22 olan bu özellik, hamilelik sayısının diyabet olasılığı üzerindeki göreceli olarak düşük etkisini yansıtmaktadır.

1. **Grafikler**

**3.1. Kabarcık Grafiği**

Analiz sürecinin bir parçası olarak, yaşlara göre glikoz değerlerini gösteren renkli bir kabarcık grafiği oluşturulmuştur. Bu grafik, veri setindeki her bireyin yaşına ve glikoz seviyesine dayalı olarak bir kabarcık olarak temsil edilmiştir. Kabarcıkların rengi, bireyin diyabet hastası olup olmadığını göstermektedir; kırmızı diyabet hastası olmayı, mavi ise diyabet hastası olmamayı simgeler.

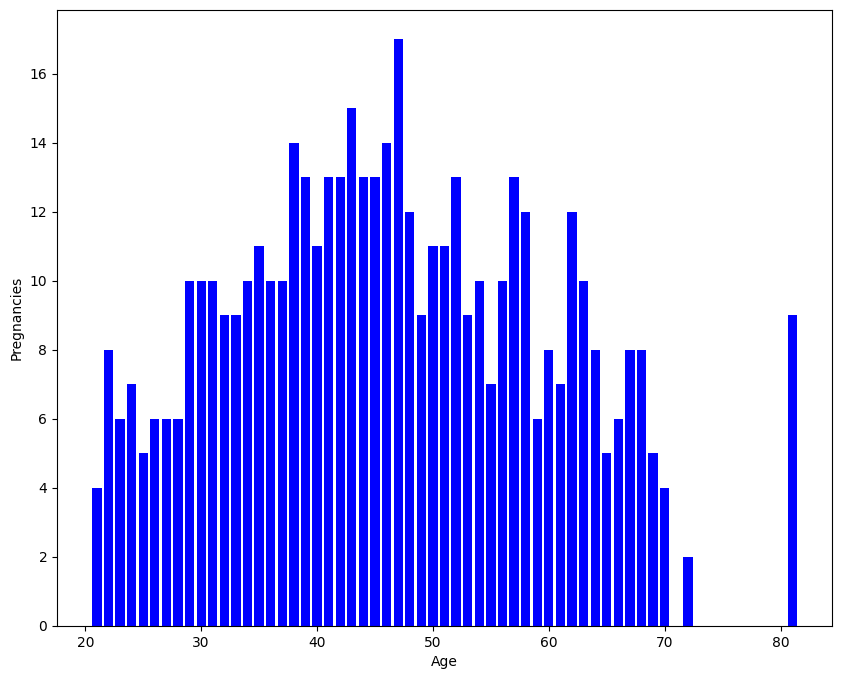
Kabarcık grafiği, yaş ile glikoz değeri arasındaki ilişkiyi görselleştirerek, diyabet olasılığını yaşa göre nasıl etkilediğini anlamamıza da yardımcı olur. Grafikte, yaşın artmasıyla birlikte glikoz seviyelerinin nasıl değiştiği ve diyabet hastalığına sahip olma olasılığının bu değişikliklerle nasıl ilişkilendiği açıkça görülmektedir. Bu görsel analiz, yaşın diyabet olasılığı üzerindeki etkisini daha iyi anlamamıza yardımcı olurken, aynı zamanda glikoz seviyelerinin diyabet riski üzerindeki önemini vurgular.



**3.2. Çubuk Grafiği**

Analiz sürecinin devamında, yaşlara göre maksimum hamilelik sayısını gösteren bir çubuk grafiği oluşturulmuştur. Bu grafik, her yaş grubundaki bireylerin maksimum hamilelik sayısını gösteren çubuklarla temsil edilmiştir. Grafik, yaşın hamilelik sayısı üzerindeki etkisini göstermektedir.

Çubuk grafiği, yaşa bağlı olarak hamilelik sayısının nasıl değiştiğini ve yaşın bu değişiklik üzerindeki etkisini görsel olarak gösterir. Grafik, yaş gruplarının hamilelik sayısını karşılaştırarak, yaşın hamilelik sayısı üzerindeki dağılımını ve yaş grupları arasındaki farklılıkları açıkça ortaya koyar. Bu analiz, yaşın hamilelik sayısına olan etkisini anlamamıza yardımcı olurken, aynı zamanda yaşa bağlı olarak hamilelik sayısındaki değişiklikleri vurgular.



**4. Model**

**4.1. Modeller**

Proje kapsamında, diyabet teşhisi koymak için beş farklı makine öğrenimi modeli kullanılmıştır. Bu modeller şunlardır:

**4.1.1. Lojistik Regresyon:**

Lojistik regresyon, bir olayın olasılığını tahmin etmek için kullanılan bir istatistiksel modeldir. Bu model, bağımsız değişkenlerle bağımlı bir değişken arasındaki ilişkiyi modellemek için kullanılır. Proje kapsamında, lojistik regresyon modeli diyabet riskini tahmin etmek için kullanılmıştır.

**4.1.2. Destek Vektör Makineleri (SVM):**

Destek vektör makineleri (SVM), sınıflandırma ve regresyon problemleri için kullanılan bir makine öğrenimi modelidir. SVM, veri noktalarını sınıflandırmak için bir karar sınırı oluşturur ve bu sınıflandırma işlemini en iyi şekilde yapacak şekilde optimize edilir. Proje kapsamında, SVM modeli diyabet teşhisi koymak için kullanılmıştır.

**4.1.3. Naive Bayes:**

Naive Bayes, olasılık teorisine dayalı bir sınıflandırma algoritmasıdır. Bu model, Bayes teoremi ve özellikler arasındaki bağımsızlık varsayımını kullanarak sınıflandırma yapar. Proje kapsamında, Naive Bayes modeli diyabet teşhisi koymak için kullanılmıştır.

**4.1.4. Karar Ağacı:**

Karar ağacı, sınıflandırma ve regresyon problemleri için kullanılan bir modeldir. Bu model, bir veri kümesindeki özelliklerin değerlerine göre kararlar alarak sonuçları tahmin eder. Proje kapsamında, karar ağacı modeli diyabet teşhisi koymak için kullanılmıştır.

**4.1.5. Yapay Sinir Ağı:**

Yapay sinir ağı, biyolojik sinir ağlarını taklit eden bir makine öğrenimi modelidir. Bu model, birbirine bağlı yapay sinir hücrelerinden oluşan katmanlar arasındaki ilişkileri öğrenir ve karmaşık veri kümelerindeki desenleri tanımlamak için kullanılır. Proje kapsamında, yapay sinir ağı modeli diyabet teşhisi koymak için kullanılmıştır.

Bu modeller, diyabet teşhisi koymak için kullanılan farklı yaklaşımları temsil eder ve her biri veri setinin farklı özelliklerini ve yapısını dikkate alarak sonuçlar üretir.

**4.2. Modellerin Eğitimi ve Test Gruplarının Oluşturulması**

Veri görselleştirmesi ve analizi yapıldıktan sonra, eğitim ve test grupları oluşturulmuştur. Veri seti eğitim ve test verisi olmak üzere iki kısma ayrılmıştır. Eğitim seti, modelin öğrenme sürecinde kullanılmış ve modelin parametreleri bu veri üzerinde ayarlanmıştır. Test seti ise modelin performansının değerlendirilmesi için kullanılmıştır.

Veri seti, %66 eğitim ve %34 test olacak şekilde rastgele bölünmüştür. Bu bölünme işlemi, her sınıftan eşit miktarda veri içeren ve veri setinin dağılımını bozmayan bir şekilde gerçekleştirilmiştir.

Eğitim setinin (%66) ve test setinin (%34) ayrı tutulmasıyla elde edilen sonuçlar, her bir model için ayrı ayrı değerlendirilmiştir. Bu değerlendirme sonuçları, her modelin eğitim ve test verisi üzerindeki performansını belirlemek için kullanılmıştır. Bu süreç, modelin doğruluğunu, hassasiyetini ve diğer performans metriklerini değerlendirmek için kullanılmıştır.

**5. Eğitimler**

Eğitimler, test seti ve verinin tamamı olmak üzere iki defa gerçekleştirilmiştir. Her bir model için eğitim sonuçları elde edilmiş ve çeşitli metriklerle değerlendirilmiştir.

**5.1. Modellerin Eğitime Göre Sonuçlarının Değerlendirilmesi**

Her eğitim sonucunda modelin 0 ile 1 arasında bir doğruluk skoru vermesi beklenir.

**5.1.1. Eğitim Verisinin Bölünmeden Kullanıldığı Durum**

Eğitim verisinin %100 olarak bölünmeden kullanıldığı durumda elde edilen sonuçlar ve yorumları aşağıdaki gibidir:

* Lojistik Regresyon Doğruluk Skoru: 0.7669856459330144
* Destek Vektör Makineleri (SVM) Doğruluk Skoru: 0.7656527682843473
* Naive Bayes Doğruluk Skoru: 0.7734962406015038
* Karar Ağacı Doğruluk Skoru: 0.6717703349282296
* Yapay Sinir Ağı Doğruluk Skoru: 0.7318694463431306

**5.1.2. Eğitim Verisinin Bölünerek Kullanıldığı Durum**

Eğitim verisinin %66'sının test verisine %34'ünün dahil edilmeden elde edilen sonuçlara göre ise:

* Lojistik Regresyon Doğruluk Skoru: 0.7823529411764707
* Destek Vektör Makineleri (SVM) Doğruluk Skoru: 0.786235294117647
* Naive Bayes Doğruluk Skoru: 0.786313725490196
* Karar Ağacı Doğruluk Skoru: 0.7173725490196078
* Yapay Sinir Ağı Doğruluk Skoru: 0.7034509803921569

Modellerin eğitiminden bu sonuçlar elde edilmiş olup eğitimler rastgelelik içerdiği için eğitim sonuçları değişiklik gösterebilir.

**5.2. Eğitim Verilerinin Kaydedilmesi**

"Eğitim" verileri, "diabetes.sav" dosyasına kaydedilmiştir. Bu dosya, eğitim sonuçlarını ve modeli daha sonra kullanmak üzere saklar.

**5.3. Eğitilmiş Makinenin Test Edilmesi**

Eğitilmiş makine, "diabetes.sav" dosyasından çekilerek test edilmiş ve sonuçlar alınmıştır. Bu testler, modelin gerçek dünya verilerinde nasıl performans gösterdiğini değerlendirmek için yapılır.

**6. Sonuç**

Bu çalışma, diyabet teşhisi koymak için makine öğrenimi modellerinin kullanılabilirliğini değerlendirmiştir. Veri seti üzerinde yapılan analizler sonucunda glikoz değeri, vücut kitle indeksi (BMI) ve yaş gibi faktörlerin diyabet üzerinde belirleyici olduğu tespit edilmiştir. Bu bilgiler doğrultusunda, lojistik regresyon, destek vektör makineleri (SVM), Naive Bayes, karar ağaçları ve yapay sinir ağları gibi çeşitli modeller eğitilmiş ve değerlendirilmiştir.

Yapılan deneyler sonucunda, en yüksek doğruluk skorlarına sahip modelin Naive Bayes olduğu belirlenmiştir. Bu model, diyabet riskini değerlendirmek için kullanılabilecek güvenilir bir araç olabilir. Ayrıca, eğitilmiş modelin "diabetes.sav" dosyasına kaydedilerek daha sonraki çalışmalarda veya uygulamalarda kullanılmak üzere saklandığı tespit edilmiştir.

Sonuç olarak, bu çalışma, makine öğrenimi tekniklerinin diyabet teşhisi koymada potansiyelini göstermiş ve diyabet hastalığının erken teşhisine yönelik önemli bir adım olmuştur.

**7. Kaynakça:**  
Veri Seti: https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database