Makine Öğrenmesi Projesi Raporu

Proje Adı: Diyabet Veri Seti Üzerinde Makine Öğrenmesi Modelleri ile Tahminleme

Bu proje, diyabet hastalığı tahminine yönelik olarak makine öğrenmesi modellerinin karşılaştırılması ve en iyi modelin belirlenmesini amaçlamaktadır. Projede hem dengeli hem dengesiz veri setleri üzerinde çalışılmış. Kullanıcı dostu bir arayüz ile model seçimi ve tahmin işlemleri kolaylaştırılmıştır.

Veri Setleri: Projede kullanılan veri setleri:

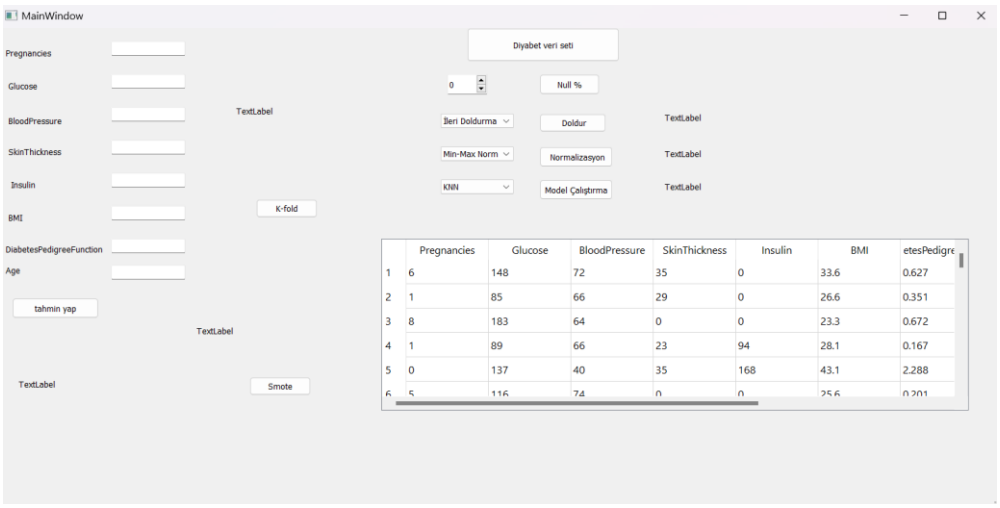
- **Orijinal Veri Seti:** diabetes.csv
- **Dengesiz Veri Seti:** diabetes_undersampled.csv

Veri seti 768 veri içermektedir. Veri setleri, Pregnancies, Glucose, BloodPressure, SkinThickness, Insulin, BMI, DiabetesPedigreeFunction ve Age değişkenlerini içermektedir. Hedef değişken "Outcome" olup, 1 diyabet varlığı, 0 diyabet yokluğunu temsil etmektedir.

Orijinal veri setinde hedef değişken "Outcome" da 0 lar 500 ,1 ler 268 tanedir. Burada 1 lerin sayısını düşürüp 100 yaparak orijinal veri setini dengesiz veri seti yaptım. Bu veri setini “diabetes_undersampled.csv” kaydettim.

Kullanıcı Arayüzü Tasarımı: Ana ekran, modern ve kullanıcı dostu bir tasarıma sahiptir. Ekran bileşenleri:

- **Veri Yükleme:** Kullanıcı, orijinal ve dengesiz veri setlerinden birini yükleyebilir.
- **ComboBox (Model Seçimi):** KNN, Karar Ağacı ve Random Forest modellerinden birini seçebilir.
- **Model Seçimi (K- fold Buttonu):** Orijinal veri seti üzerinde seçilen modelle göre k-fold uygulanması
- **Model Eğit Butonu:** Seçilen model eğitilir ve model başarım metrikleri gösterilir.
- **Tahmin Yap Butonu:** Eğitilen model ile tahmin yapılır ve predict_proba ve predict sonuçları ekranda gösterilir.



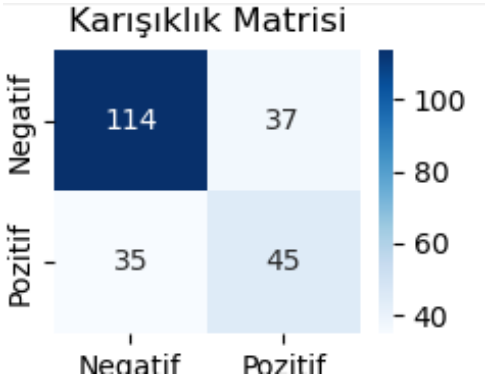
Kullanılan Modeller: Projede kullanılan modeller:

- **KNN**
- **Decision Tree Classifier (Karar Ağacı)**
- **Random Forest Classifier**
- **K-Fold Çapraz Doğrulama**

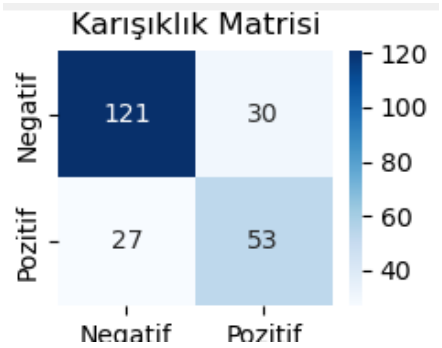
1)Orjinal Veri Seti Üzerinde Yapılan Çalışmalar

ComboBox dan seçile 3 metot ile model eğitimi gerçekleşir. Bu eğitilen modelin doğruluk (Accuracy), duyarlılık (Recall), özgüllük (Specificity) ,F1 Skoru ve karmaşıklık matrisi sonuçları kullanıcıya verilmiştir.

1)Knn Uygulanması ve sonuçları :

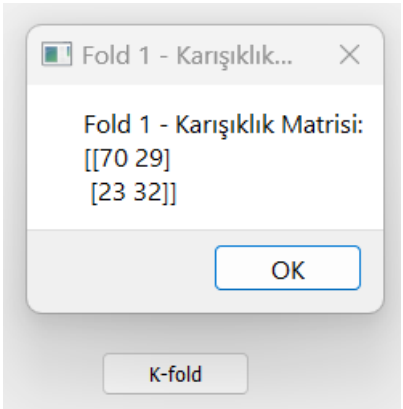


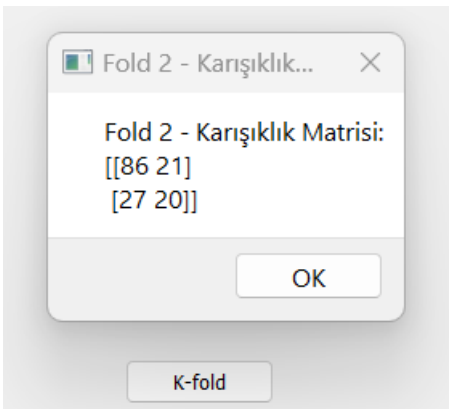
2)Random Forest Modeli ve Sonucu:

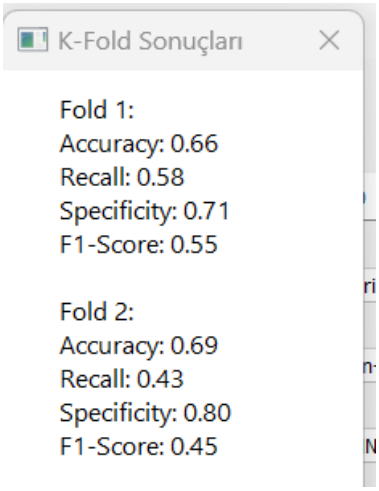


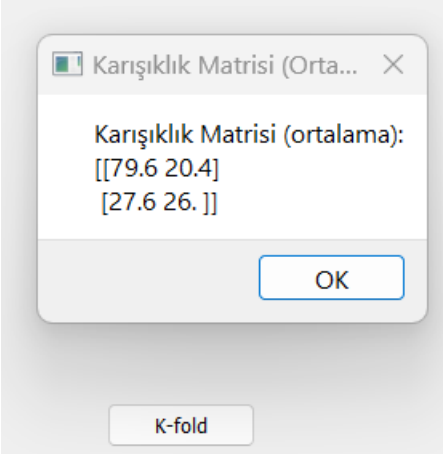
Orijinal veri seti üzerinde k-fold uygulanabilir. K-Fold Çapraz Doğrulama, veri setini k eşit parçaya böler ve her seferinde bir parçasını test veri seti olarak ayırırken kalan k-1 parçayı model eğitimi için kullanır.

Her bir k (5) değeri için ayrı ayrı başarı metrikleri ve karmaşıklık matrisi oluşturur.



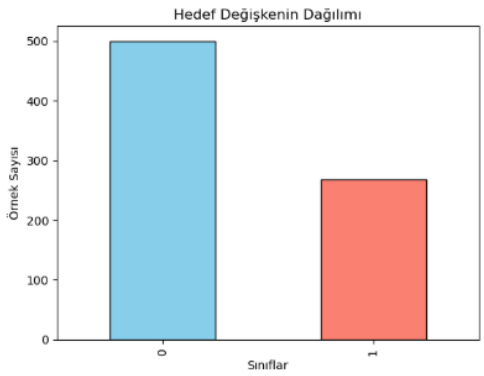




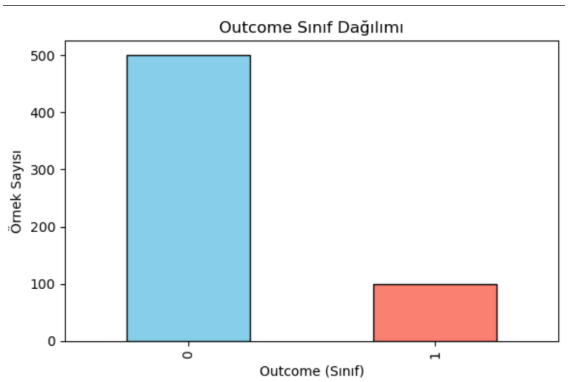


2)Dengesiz Veri seti üzerinde yapılan çalışmalar

Orijinal veri seti üzerinde yapılan işlem sonucu elde edilen dengesiz veri seti üzerinde ComboBox üzerinde seçilen 3 model ile çalışılabilir. Ve bu modellerin metrikleri ve karışıklık matrisi kullanıcıya verilir.



Şekil 1-Orijinal veri seti hedef değişken

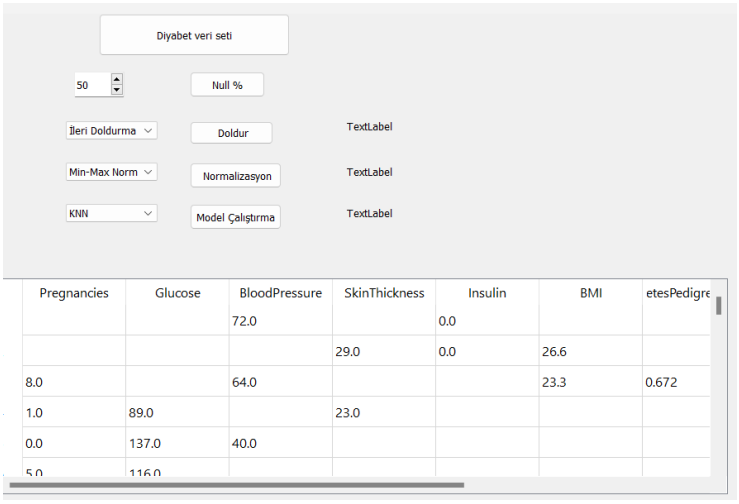


Şekil 2-Dengesiz veri setindeki hedef değişken

3)Gürültülü Veri seti ile Çalışma

Gürültülü veriler, makine öğrenmesi modellerinin performansını olumsuz etkileyebilir ve yanlış tahminlere yol açabilir. Gürültüyü azaltmak için veri temizleme ve filtreleme teknikleri uygulanır.

Kullanıcı orijinal veri seti üzerinde istediği yüzde de null değer atayarak gürültülü veri seti elde eder. Bu gürültülü veri seti üzerinde comboBoxdan aldığı yönteme göre ileri ile doldurma, geri ile doldurma ve ortalama ile doldurma seçenekleri ile gürültülü olan veri setini düzeltir. Düzeltilen veri seti üzerinde knn, random forest ve karar ağcı modellerinden biri seçilerek model eğitilir ve bunun sonuçları kullanıcıya verilir. (Başarı metrikleri ve karmaşıklık matrisi)

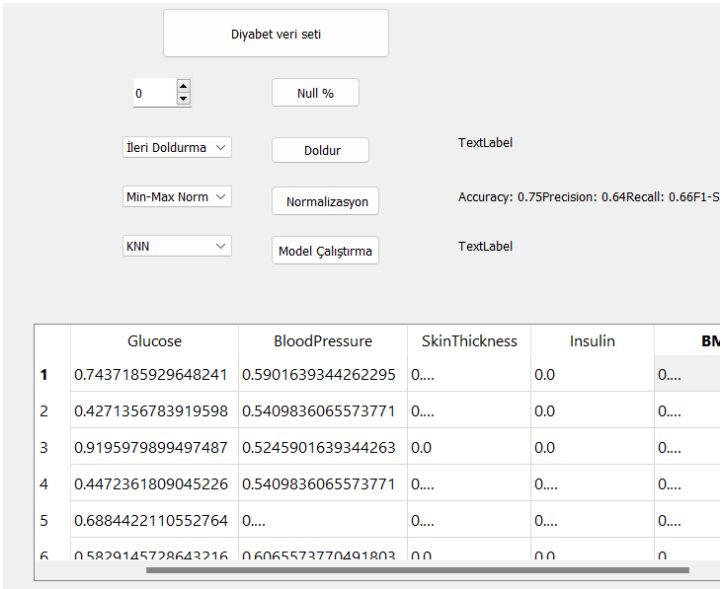


4)Normalizasyon Deneyi

Normalizasyonun amacı, veri setindeki değişkenlerin farklı ölçeklerde olmasının modellerin performansını olumsuz etkilemesini önlemektir. Normalizasyon, tüm değişkenleri aynı ölçek aralığına getirerek (genellikle [0,1] veya z-skor standardizasyonu ile) modelin daha dengeli ve doğru tahminler yapmasını sağlar.

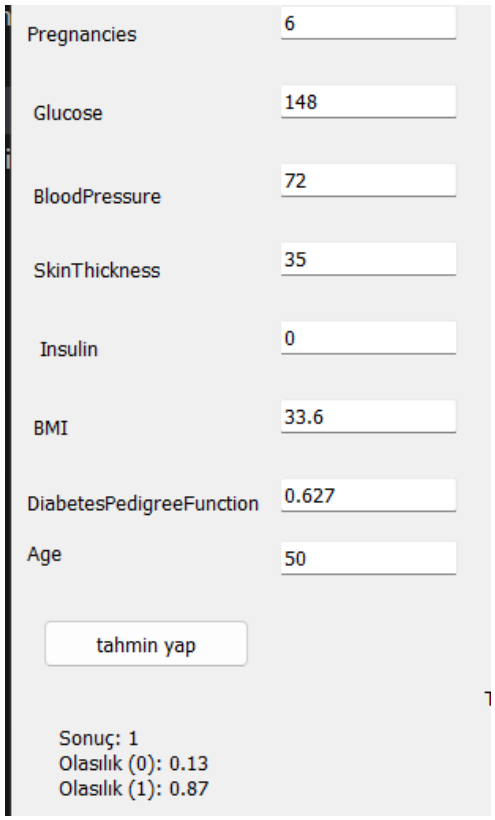
Orijinal veri setindeki hedef değişken “Outcome” sınıfı dışında kalan değişenler üzerinde normalizasyon uygulanır. İki çeşit normalizasyon uygulanabilir. Min-max normalizasyonu ve z-skor standardizasyonu uygulanabilir.

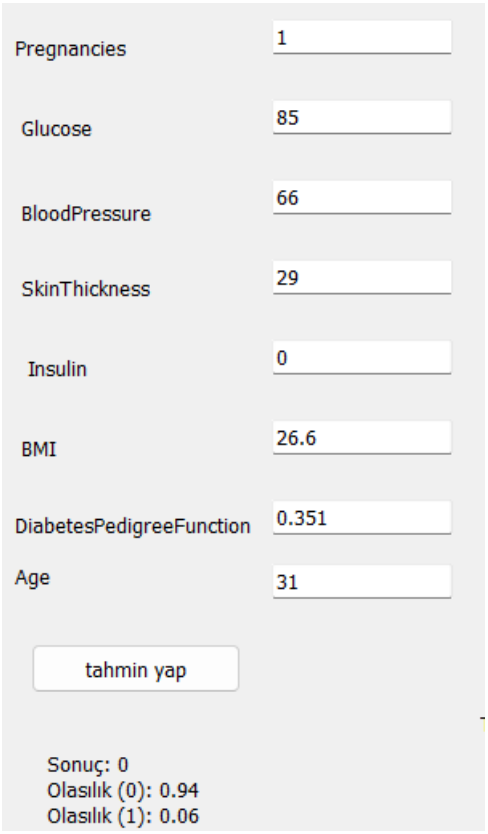
Kullanıcı seçtiği üç modelden biri ile eğitim gerçekleştirebilir ve bunun başarı metrikleri ve karmaşıklık matrisi kullanıcıya verilir.



5)Kullanıcının Girdiği değerlere Göre Hedef sınıfı Tahmin Etme

Pregnancies, Glucose, BloodPressure, SkinThickness, Insulin, BMI, DiabetesPedigreeFunction ve Age değişkenlerini kullanıcıdan alınır ve Outcome sınıfı tahmin edilir.





Tahmin (predict metodu): Model, kullanıcıdan alınan özellikler ile predict metodunu kullanarak hedef değişkenin tahminini gerçekleştirdi. Örneğin, kullanıcının girdiği verilere göre model hedef değişkenin 0 mı yoksa 1 mi olacağını tahmin etti.

Olasılık (predict_proba metodu): predict_proba metodu ise her iki sınıf için (0 ve 1) olasılıkları hesapladı. Bu, modelin tahminin doğruluğu hakkında daha fazla bilgi sağlar; örneğin, modelin tahmin ettiği 1 sınıfının olasılığı %75, 0 sınıfının olasılığı ise %25 olabilir. Bu olasılıklar, modelin karar verme sürecine dair güven seviyesini yansıtır.