Veri Ön İşleme Süreçlerinin Geliştirilmesi

Veri işleme ve dengeleme süreci, ham veriyi analiz ve modelleme için uygun hale getirmek amacıyla gerçekleştirilen bir dizi adımı içerir. Bu süreçte, “output.csv” dosyasında bulunan veriler işlenmiş, öznitelik dönüşümleri yapılmış ve sınıf dengesizliği SMOTE yöntemi ile giderilmiştir. İşlemler sonucunda, veri setinin analitik modelleme için daha kullanışlı ve dengeli bir formata dönüştürülmesi sağlanmıştır.

1. Veri Yükleme

İlk olarak ”output.csv” dosyası yüklenmiştir. Bu dosya, hemoglobin (Hb) seviyeleri, renk pikselleri ve diğer ilgili özellikleri içeren bir veri setidir.

1. Veri Ön İşleme

2.1 Cinsiyet (Gender) Dönüşümü

Gender sütunundaki harfler (“F”,” M”,”F “,”M ”) sayısal değerlere (0, 1) dönüştürülmüştür.

Amaç: Cinsiyet bilgisini sayısal formata dönüştürmek, veri analizi ve modelleme süreçlerinde kullanılabilir hale getirmek.

2.2 Anemi (Anaemic) Dönüşümü

Anaemic sütunu, etiketleme (Label Encoding) yöntemi ile sayısal değerlere dönüştürülmüş ve ardından One-Hot Encoding yöntemi ile ikili sütunlara ayrılmıştır.

Amaç: Kategorik verileri modelleme sürecinde kullanılabilir hale getirmek ve her kategoriye özel sütunlar oluşturmak.

1. Öznitelik Ölçekleme ve Hesaplamalar

3.1 Ölçekleme

'%Red Pixel', '%Green pixel', '%Blue pixel', ve 'Hb' sütunları Min-Max Scaling yöntemi ile [0, 1] aralığına dönüştürülmüştür.

Amaç: Özelliklerin eşit derecede önemli olmasını sağlamak ve modelleme süreçlerinde daha iyi performans elde etmek.

3.2 Yeni Özellikler

Ortalama Piksel: Renklerin ortalama değerini hesaplama.

Renk Farkları: Renkler arasındaki farkları hesaplayarak mutlak değerler alınmıştır.

Log Hb: Hemoglobin seviyelerinin logaritması hesaplanmış, sıfır değerleri minumum pozitif değere dönüştürülmüştür.

Ek Özellikler: Red-Blue Product, Hb to Red Ratio hesaplanmış, çeşitli ilişkiler ve oranlar değerlendirilmiştir.

1. Risk Faktörü Hesaplama

Hb ve '%Red Pixel' değerleri kullanılarak risk faktörü hesaplanmış, “0”, “0.5” ve “1” değerleri atanmıştır.

Amaç: Klinik riskleri modellemek ve hastalık tespiti için önemli bir öznitelik oluşturmak.

1. Veri Dengeleme

5.1 SMOTE Uygulaması

İşlem: SMOTE (Synthetic Minority Over-sampling Technique) yöntemi kullanılarak azınlık sınıfların örnekleri artırılmıştır.

Amaç: Veri setindeki sınıf dengesizliğini gidermek ve model performansını artırmak.

5.2 Risk ve Anaemi Sütunlarının Düzenlenmesi

İşlem: SMOTE uygulaması sonrası risk ve anaemi sütunlarında oluşan beklenmeyen değerler düzeltilmiştir. “Anaemic\_1” sütunu düzenlenmiş ve Risk sütunu “0”, “0.5” ve “1” değerlerini içerecek şekilde ayarlanmıştır.

6. Sonuç ve Kaydetme

Veri seti işlenmiş , öznitelikler dönüştürülmüş ve dengeleme işlemleri gerçekleştirilmiştir. Sonuç olarak “Balanced\_Processed\_Data.csv“ dosyasında güncellenmiş ve temizlenmiş veri seti elde edilmiştir. Bu veri seti, analiz ve modelleme süreçleri için uygun hale getirilmiştir.