FETAL SAĞLIK VERİ SETİ İLE SINIFLANDIRMA ANALİZİ

1. GİRİŞ (Problem Tanımı)

Gebelik sürecinde fetüsün sağlık durumunun erken ve doğru şekilde belirlenmesi, hem anne hem de bebek sağlığı açısından büyük önem taşımaktadır. Klinik ortamda elde edilen ölçümlerin doğru yorumlanması, olası risklerin önceden tespit edilmesini sağlar. Ancak bu verilerin manuel olarak değerlendirilmesi zaman alıcı ve hata payı yüksek bir süreçtir.

Bu projede, fetal sağlık verileri kullanılarak fetüsün sağlık durumunun Normal, Suspect veya Pathological olarak sınıflandırılması amaçlanmıştır. Problem, çok sınıflı bir sınıflandırma problemi olarak ele alınmış ve farklı makine öğrenmesi algoritmaları kullanılarak çözümlenmiştir.

2. YÖNTEM

Projede kullanılan veri seti, fetüse ait kalp atım hızı, kasılma sayıları, kısa ve uzun dönem varyasyon ölçümleri ile histogram tabanlı özellikleri içermektedir. Toplamda 21 adet bağımsız değişken ve 1 adet hedef değişken bulunmaktadır.
Hedef değişken olan fetal_health üç sınıftan oluşmaktadır:

1: Normal

2: Suspect

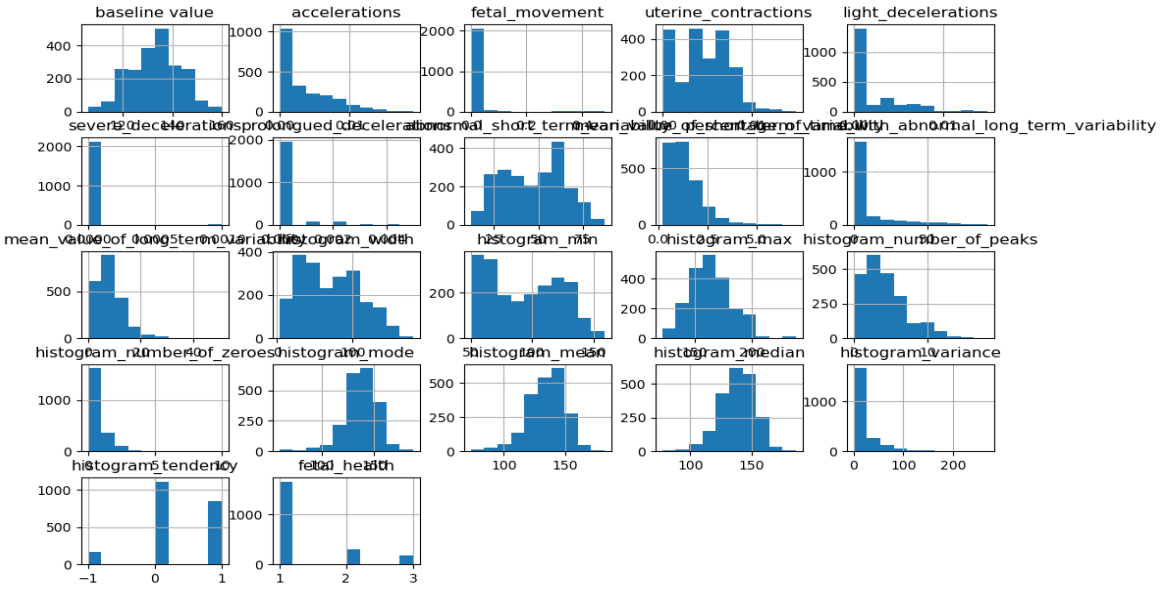
3: Pathological

Veri seti üzerinde yapılan eksik veri analizinde herhangi bir eksik gözlem bulunmadığı tespit edilmiştir.

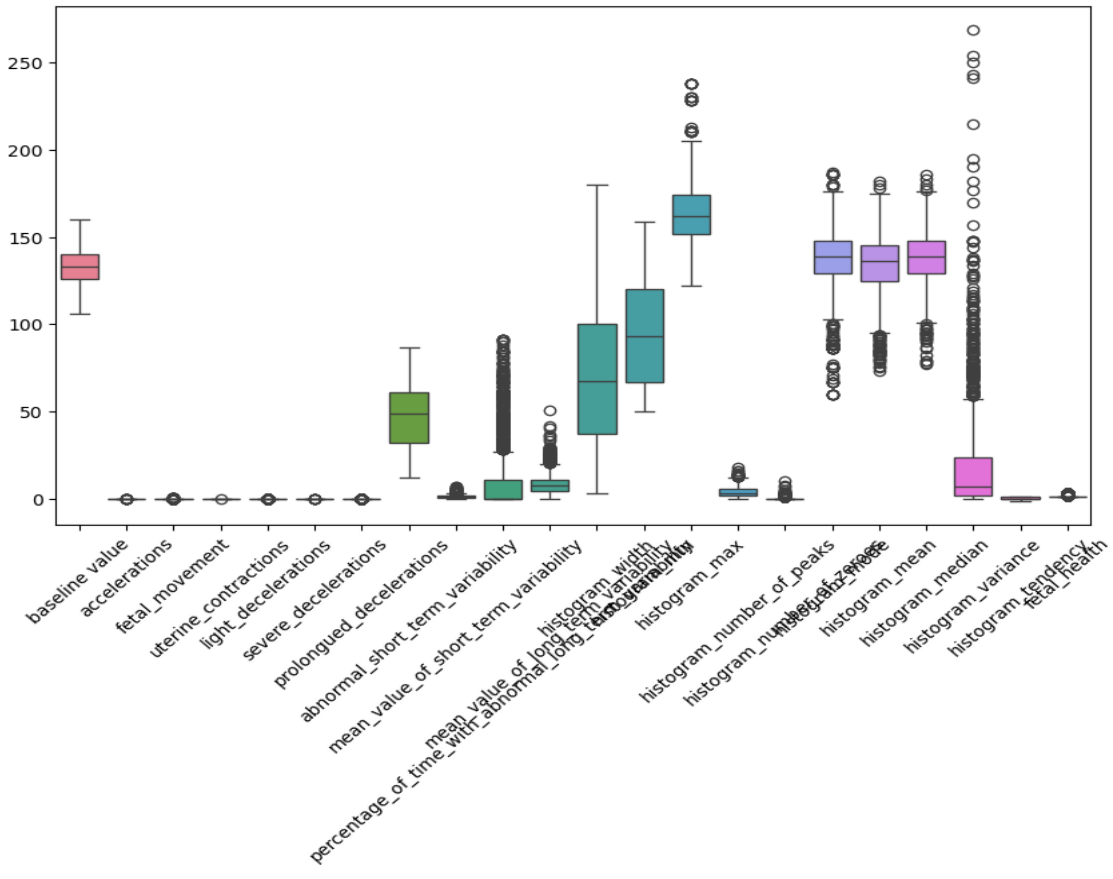
2.2 Keşifçi Veri Analizi (EDA)

Veri seti üzerinde keşifçi veri analizi yapılmış, değişkenlerin dağılımları histogram ve boxplot grafiklerle incelenmiştir. Ayrıca değişkenler arasındaki ilişkileri görmek amacıyla korelasyon matrisi (heatmap) oluşturulmuştur.
Bazı değişkenler arasında anlamlı korelasyonlar gözlemlenmiş, bu durum özellikle sınıflandırma algoritmalarının performansını etkilemiştir.

FETAL SAĞLIK VERİ SETİ İLE SINIFLANDIRMA ANALİZİ



Şekil 1. Veri setindeki değişkenlerin dağılımlarını gösteren histogram grafikleri.



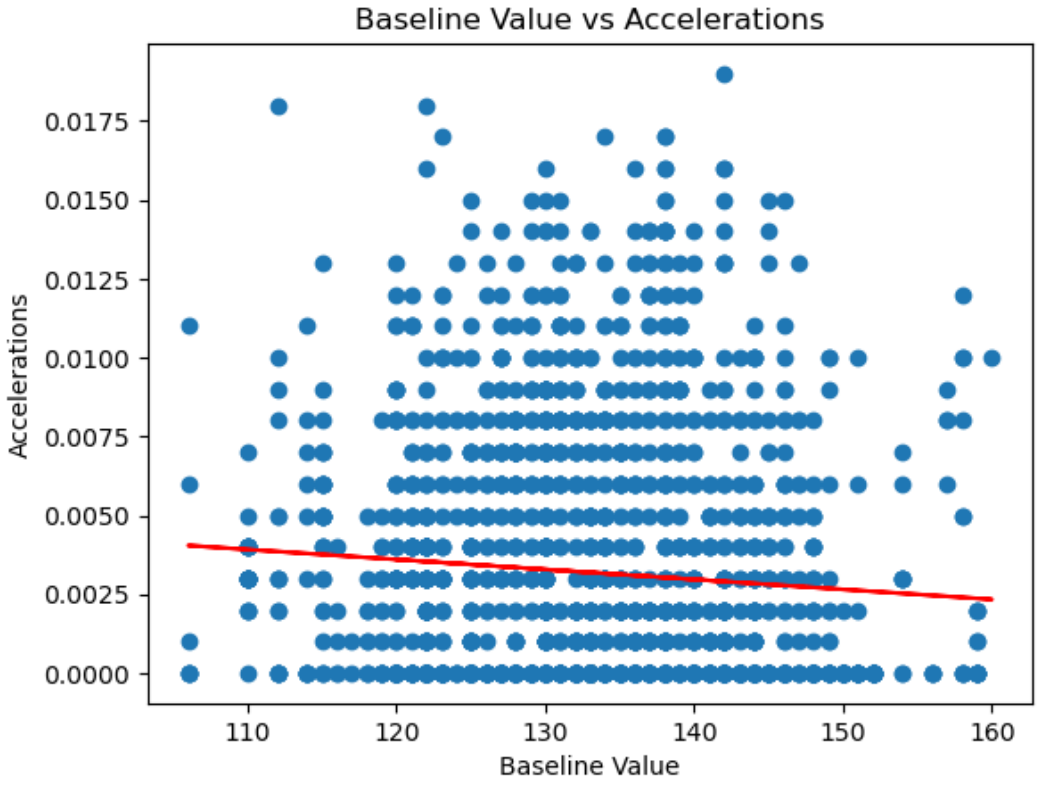
Şekil . eğişkenlere ait a kırı değerlerin o lot ile gösterimi.

Şekil 1 ve Şekil 2 incelendiğinde, bazı değişkenlerin (örneğin fetal_movement ve severe_decelerations) sağa çarpık bir dağılım sergilediği, ayrıca bazı özelliklerde belirgin aykırı değerlerin bulunduğu gözlemlenmiştir.

2.3 Lineer İlişki ve Regresyon Analizi

Değişkenler arasındaki olası lineer ilişkileri incelemek amacıyla iki sayısal değişken arasında lineer regresyon analizi yapılmıştır. Bu analiz kapsamında scatter plot grafiği oluşturulmuş ve regresyon doğrusu çizdirilmiştir.

Elde edilen sonuçlara göre seçilen değişkenler arasında güçlü bir lineer ilişki bulunmadığı görülmüştür. Regresyon modeline ait R^2 değeri düşük çıkmış olup bu durum, değişkenler arasındaki ilişkinin zayıf olduğunu göstermektedir. Bu analiz, veri setindeki ilişkilerin büyük ölçüde doğrusal olmayan yapıda olduğunu ortaya koymuştur.



Şekil 3. Baseline Value ile Accelerations arasındaki ilişki ve lineer regresyon doğrusu.

2.4 Kullanılan Sınıflandırma Modelleri

Bu çalışmada fetal sağlık durumunun tahmin edilmesi amacıyla dört farklı sınıflandırma algoritması kullanılmıştır. Bu algoritmalar sırasıyla Logistic Regression, K-Nearest Neighbors (KNN), Support Vector Machine (SVM) ve Decision Tree'dir.

KNN algoritmasında en uygun komşu sayısını belirlemek için farklı k değerleri denenmiş ve en iyi performansı veren değer seçilmiştir. Modellerin değerlendirilmesinde Accuracy, Precision, Recall ve F1-Score metrikleri kullanılmıştır. Ayrıca sınıflandırma sonuçları Confusion Matrix yardımıyla detaylı olarak analiz edilmiştir.

3. BULGULAR

3.1 Model Performans Sonuçları

Uygulanan sınıflandırma modelleri karşılaştırıldığında tüm modellerin yüksek doğruluk oranlarına ulaştığı görülmüştür. Ancak sınıf bazlı performanslar incelendiğinde modeller arasında belirgin farklar olduğu tespit edilmiştir.

Decision Tree modeli, özellikle Pathological sınıfındaki örnekleri doğru tahmin etme konusunda diğer modellere kıyasla daha başarılı sonuçlar vermiştir. Bu modelin hem Recall hem de F1-Score değerlerinin yüksek olması, sağlık verileri açısından güvenilir bir performans sunduğunu göstermektedir.

3.2 Model Karşılaştırması

Modeller genel performans açısından karşılaştırıldığında Decision Tree algoritmasının en dengeli sonuçları verdiği görülmüştür. KNN ve SVM modelleri de başarılı tahminler yapmış olsa da Decision Tree modelinin sınıf bazlı başarısı daha tutarlı bulunmuştur.

Logistic Regression modeli ise özellikle sınıflar arasındaki karmaşık ilişkileri yakalama konusunda diğer modellere göre daha sınırlı kalmıştır. Bu nedenle nihai model olarak Decision Tree algoritması tercih edilmiştir.

4. SONUC

Bu projede fetal sağlık verileri kullanılarak farklı makine öğrenmesi algoritmaları ile sınıflandırma analizi gerçekleştirilmiştir. Yapılan analizler sonucunda makine öğrenmesi yöntemlerinin sağlık alanında etkili bir karar destek aracı olarak kullanılabileceği görülmüştür.

Modeller arasında yapılan karşılaştırma sonucunda Decision Tree algoritmasının hem genel doğruluk hem de sınıf bazlı performans açısından en başarılı model olduğu belirlenmiştir. Özellikle patolojik vakaların doğru şekilde tespit edilmesi, bu modelin tercih edilmesinde önemli bir etken olmuştur.

Elde edilen sonuçlar, benzer veri setleri üzerinde yapılacak çalışmalar için yol gösterici niteliktedir.