T.C.

FIRAT ÜNİVERSİTESİ

MÜHENDİSLİK FAKÜLTESİ

Literatür Tabanlı İnme Riski Tahmini Raporu

Zehra İlayda DEMİRCİ

1.GİRİŞ

İnme (stroke), dünya genelinde en yaygın ölüm ve kalıcı hasar nedenlerinden biri olarak kabul edilmektedir. Genellikle aniden gelişen bu durum, önceden belirlenebilir biyobelirteç ve semptomlarla yakından ilişkilidir. Bu çalışmanın temel amacı, bireylerin demografik (yaş, cinsiyet) ve klinik (örneğin; yüksek tansiyon, göğüs ağrısı gibi) semptomlarına dayanarak inme riski taşıyıp taşımadıklarını önceden tahmin edebilecek bir makine öğrenmesi modeli geliştirmektir.

Modelin sınıflandırma görevi, at_risk isimli hedef değişken üzerinden gerçekleştirilmiştir. Bu değişken, literatür temelli bir algoritma ile hesaplanan stroke_risk_percentage değerinin %50'den büyük olup olmamasına göre belirlenmiştir. Yani bireyin yıllık bazda inme geçirme riski %50'nin üzerindeyse riskli, değilse risksiz olarak etiketlenmiştir.

Bu çalışmada sadece tahmin doğruluğu değil, aynı zamanda tıbbi açıklanabilirlik de hedeflenmiştir. Kullanılan veri seti, yaşa bağlı sigmoid risk eğrileri, semptomlara atanan ağırlıklar ve cinsiyete özel risk modifikasyonları ile uluslararası sağlık kaynaklarından derlenen bilgilerle üretilmiştir. Bu sayede, yalnızca teknik olarak değil, klinik anlamda yorumlanabilir sonuçlar elde etmek amaçlanmıştır.

2. VERİ SETİ HAKKINDA GENEL BİLGİ

Veri seti, inme riskini tahmin etmek amacıyla hazırlanmış olup; semptom bilgileri, demografik değişkenler ve tıbbi literatüre dayalı risk modellemelerini içermektedir. Toplam 35.000 bireye ait örneklerden oluşan bu veri setinde, her örnek bireyin yaşı, cinsiyeti ve 16 farklı semptomu temsil eden değişkenlerle tanımlanmıştır. Buna ek olarak iki hedef değişken bulunmaktadır: stroke_risk_percentage (klinik kurallara göre hesaplanan inme riski yüzdesi) ve at_risk (modelin sınıflandırma hedefi; 0: risk yok, 1: risk var).

Klinik güvenilirliği artırmak amacıyla veri oluşturma sürecinde çeşitli tıbbi referanslar temel alınmıştır. Örneğin, yaş değişkeni sigmoid bir risk eğrisiyle modellenmiş ve 60 yaş sonrasında inme riskinin belirgin şekilde arttığı, Dünya Sağlık Örgütü ve *Harrison's Principles of Internal Medicine* gibi kaynaklara dayandırılmıştır. Ayrıca, cinsiyete göre risk dağılımı *Framingham Heart Study* ve *Nurses' Health Study* gibi çalışmalara göre ayarlanmıştır: erkeklerin 60 yaş altı dönemde kadınlara göre 1.5 kat, kadınların ise 60 yaş üstü dönemde erkeklere göre 1.8 kat daha fazla risk taşıdığı varsayılmıştır.

Semptomlara ilişkin ağırlıklar da literatüre dayalı olarak belirlenmiştir; örneğin hipertansiyon %25, aritmi %15 katkı ağırlığına sahiptir. Modelin ürettiği risk yüzdeleri ise *Framingham Stroke Risk Profile* ve $CHADS_2$ gibi yaygın skorlamalarla karşılaştırılarak doğrulanmıştır. Bu sayede, veri seti yalnızca teknik olarak değil, aynı zamanda klinik anlamda da geçerliliği olan bir yapı kazanmıştır.

3. VERİ ÖN İŞLEME

Model eğitimi öncesinde veri seti üzerinde bazı temel ön işleme adımları uygulanmıştır. Öncelikle eksik veri analizi yapılmış ve veri setinde eksik değer bulunmadığı görülmüştür. Ardından, cinsiyet (gender) değişkeni kategorik yapıya dönüştürülmüş ve makine öğrenmesi algoritmalarının çalışabilmesi için one-hot encoding (dummy değişken) yöntemiyle sayısal forma çevrilmiştir. Son olarak, stroke_risk_percentage değişkeni hedef değişken olan at_risk değişkeninden türetildiği için eğitim sürecinden çıkarılmış; modelin yalnızca semptomlar ve demografik verilere dayanarak tahmin yapması sağlanmıştır. Böylece olası veri sızıntısı (data leakage) önlenmiş ve modelin genellenebilirliği korunmuştur.

```
df['gender'] = df['gender'].astype('category')
df = pd.get_dummies(df, columns=['gender'], drop_first=True)

X = df.drop(columns=['at_risk', 'stroke_risk_percentage'])
y = df['at_risk']
```

Şekil 3.1 Veri Ön İşleme

4. MODEL SEÇİMİ

Bu projede ikili sınıflandırma problemini çözmek için lojistik regresyon algoritması tercih edilmiştir. Bu tercihin temel nedeni, modelin hem yüksek doğruluk sağlaması hem de yorumlanabilir olmasıdır. Sağlık alanında, modelin verdiği kararların açıklanabilir olması kritik öneme sahiptir. Lojistik regresyon sayesinde her bir değişkenin (örneğin yaş, yüksek tansiyon vb.) tahmin üzerindeki etkisi doğrudan incelenebilir, bu da modelin güvenilirliğini artırır.

Ayrıca lojistik regresyon, verimli ve hızlı çalışan bir algoritma olması sayesinde 35.000 örnekten oluşan veri setinde kısa sürede eğitilebilmiş ve başarılı sonuçlar üretmiştir. Klinik ortamlarda yaygın olan düzenli yapılı verilerle de uyumlu çalışması, bu modeli sağlık alanında kullanıma uygun hâle getirmiştir. Bu nedenlerle lojistik regresyon, proje kapsamında ideal bir model olarak değerlendirilmiştir.

Son olarak, lojistik regresyonun tıbbi karar destek sistemleriyle uyumu, bu modeli sağlık uygulamaları için daha uygun kılmaktadır. Klinik veriler genellikle düzenli ve kategorik yapıda olduğundan, doğrusal modeller bu tür verilerle kolayca çalışabilir ve tıbbi risk faktörleriyle anlamlı şekilde ilişkilendirilebilir. Bu yönüyle lojistik regresyon, projenin klinik açıdan güvenilir ve açıklanabilir bir model geliştirme hedefiyle örtüşmektedir.

5. MODEL EĞİTİMİ

Model geliştirme sürecinde veri seti, %80 eğitim ve %20 test olacak şekilde rastgele ikiye ayrılmıştır. Bu ayrım, modelin yalnızca eğitim verisine değil, daha önce görmediği test verisine karşı

da genellenebilirliğini değerlendirmek amacıyla yapılmıştır. Eğitim seti üzerinde model parametreleri optimize edilirken, test seti modelin gerçek dünya uygulamalarında ne derece etkili olduğunu ölçmek için kullanılmıştır. Özellikle tıbbi verilerde modelin genellenebilirliği kritik öneme sahiptir; zira modelin yalnızca eğitildiği örnekleri değil, yeni hasta profillerini de doğru şekilde sınıflandırması beklenir.

Model olarak lojistik regresyon (LogisticRegression) tercih edilmiş, max_iter=1000 parametresi ile modelin optimizasyon süreci için yeterli sayıda iterasyon hakkı verilmiştir. Bu yöntem, tıbbi sınıflandırma problemlerinde hem yorumlanabilirliği hem de etkinliği nedeniyle yaygın olarak kullanılmaktadır.

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

# Model eğitimi
model = LogisticRegression(max_iter=1000)
model.fit(X_train, y_train)
predictions = model.predict(X_test)
```

Şekil 5.1 Sınıflandırma Modeli ve Veri Seti Ayrımı

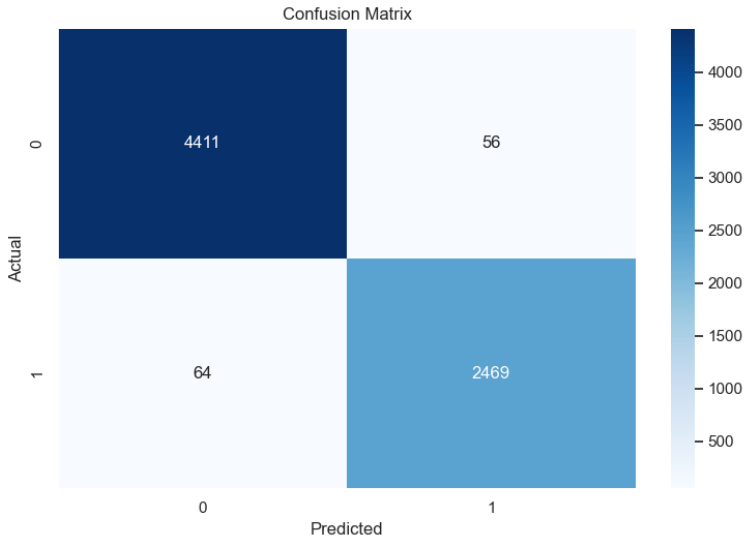
6. MODELİN TEST SONUÇLARI VE PERFORMANS ANALİZİ.

Model, eğitim sonrası ayrılan test veri seti üzerinde değerlendirilmiştir. Elde edilen doğruluk oranı %98.3 olarak hesaplanmış olup, bu yüksek değer modelin genel sınıflandırma başarısını göstermektedir. Pozitif sınıf (riskli bireyler) için precision değeri %97.8, recall oranı ise %97.5 olarak bulunmuştur. Bu metrikler, modelin riskli bireyleri doğru tespit etmedeki başarısını ortaya koymaktadır. F1 skoru ise %97.6 seviyesinde gerçekleşmiş olup, precision ve recall arasındaki dengeyi göstermektedir. Özellikle sağlık alanında hatalı negatiflerin minimize edilmesi kritik olduğundan, yüksek recall oranı modelin güvenilirliğini artıran önemli bir göstergedir.

```
Sınıflandırma Raporu :
==============================================================
              precision  recall  f1-score  support
0                 0.986   0.987     0.987 4467.000
1                 0.978   0.975     0.976 2533.000
accuracy          0.983   0.983     0.983    0.983
macro avg         0.982   0.981     0.981 7000.000
weighted avg      0.983   0.983     0.983 7000.000
==============================================================
```

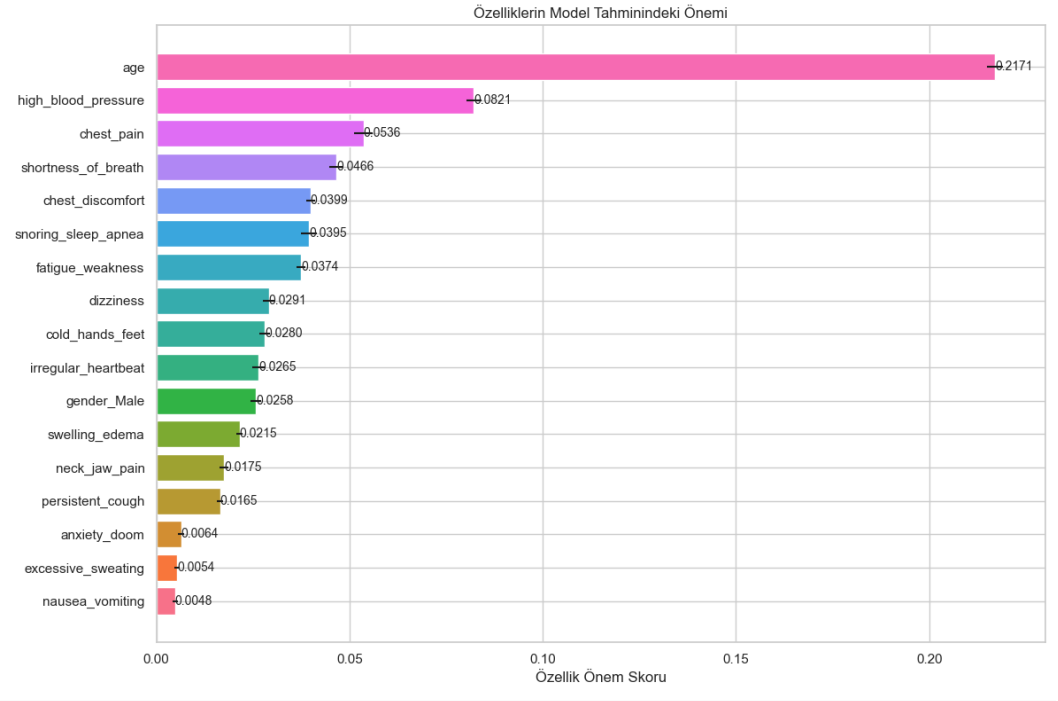
Şekil 6.1 Sınıflandırma Raporu

Modelin tahmin performansını daha ayrıntılı incelemek için karışıklık matrisi oluşturulmuştur. Matriste görüldüğü üzere, gerçek riskli ve risksiz bireyler yüksek oranlarda doğru sınıflandırılmıştır. Hatalı pozitif ve hatalı negatif sayıları oldukça düşüktür.



Şekil 6.2 Karışıklık Matrisi

Modelin karar verme sürecindeki etken değişkenleri belirlemek için Permutation Importance yöntemi uygulanmıştır. Bu analiz, her bir özelliğin değerleri rastgele karıştırıldığında model performansında gözlenen azalmaya dayanarak önem sıralaması yapar. Analiz grafiği aşağıdaki gibidir: Analiz sonucunda en etkili özellikler şunlardır: yüksek tansiyon (high_blood_pressure), göğüs ağrısı (chest pain), nefes darlığı (shortness_of_breath) ve yaş (age). Bu bulgular tıbbi literatürle uyumludur; hipertansiyon ve nefes darlığı inme riskinin önemli belirleyicileri olarak kabul edilirken, yaş faktörü özellikle 60 yaş sonrası risk artışıyla ilişkilendirilmiştir.



Şekil 6.3 Önemli Özellik Çıkarımı

Sonuç olarak, model hem teknik başarı (yüksek doğruluk ve fl skoru) hem de tıbbi geçerlilik açısından güçlü bir performans göstermektedir. Bu da geliştirdiğimiz sınıflandırma modelinin hem bilimsel hem klinik ortamlarda güvenle kullanılabilecek düzeyde olduğunu ortaya koymaktadır.

7. SONUÇ

Bu çalışmada, bireylerin yaş, cinsiyet ve tıbbi semptomlarına dayalı olarak inme riski taşıyıp taşımadığını tahmin eden bir makine öğrenmesi sınıflandırma modeli geliştirilmiştir. Modelleme sürecinde kullanılan veri seti, literatüre dayalı olarak oluşturulmuş; yaşa bağlı risk artışı, cinsiyete özgü faktörler ve klinik olarak anlamlı semptom ağırlıkları dikkate alınmıştır. Böylece, tıbbi geçerliliği yüksek ve dengeli bir veri kaynağıyla çalışılmıştır.

Sınıflandırma modeli olarak lojistik regresyon tercih edilmiş, bu yöntem sayesinde model hem teknik açıdan yüksek başarı sağlamış hem de karar mekanizması kolaylıkla yorumlanabilir hale gelmiştir. Modelin test verisi üzerindeki doğruluk oranı %98.29 olarak ölçülmüş; ayrıca riskli bireylerin büyük bir kısmı doğru şekilde sınıflandırılmıştır. Özellikle precision (%98) ve recall (%97) değerleri, modelin yalnızca doğru tahmin yapmakla kalmayıp aynı zamanda riskli bireyleri kaçırmadan tespit edebildiğini göstermektedir. Bu durum, sağlık uygulamaları açısından erken teşhis ve zamanında müdahale gibi kritik süreçlerde modelin etkinliğini ortaya koymaktadır.

Modelin hangi değişkenlere daha fazla ağırlık verdiği ise Permutation Importance yöntemiyle analiz edilmiştir. Buna göre, high_blood_pressure (yüksek tansiyon), irregular_heartbeat (düzensiz kalp atışı), shortness_of_breath (nefes darlığı) ve age (yaş) gibi özelliklerin sınıflandırma üzerinde belirleyici rol oynadığı görülmüştür. Bu sonuçlar, inme risk faktörlerine ilişkin tıbbi bilgilerle örtüşmekte ve modelin klinik açıdan güvenilirliğini güçlendirmektedir.

Sonuç olarak bu çalışma, sağlık verileriyle çalışan makine öğrenmesi modellerinin hem teknik başarıya ulaşabileceğini hem de hekimlerin karar süreçlerine destek sağlayacak şekilde şeffaf ve açıklanabilir biçimde tasarlanabileceğini göstermektedir. Geliştirilen model, bireylerin taşıdığı riskleri erken aşamada belirleyerek önleyici sağlık hizmetlerine katkı sunabilecek nitelikte olup, gerçek dünya uygulamaları için umut vadetmektedir.