İSTANBUL TOPKAPI ÜNİVERSİTESİ

MÜHENDİSLİK FAKÜLTESİ

BİLGİSAYAR MÜHENDİSLİĞİ

Ders/Dönem: FET309 - Görsel Programlama / 2025-2026 Güz Dönemi

Proje Başlığı: Makine Öğrenmesi Algoritmaları ile İnme Riski Tahmini ve Karşılaştırmalı Analizi

Öğrenci Ad Soyad: Umut Torun

Öğrenci Numara: 23040101063

Öğrenci E-Posta: umuttorun@stu.topkapi.edu.tr

Öğrenci İmza:

1. PROBLEM TANIMI & MOTİVASYON

1.1. İş/Bilimsel Soru

Dünya Sağlık Örgütü verilerine göre inme, dünya genelinde ölümlerin ve ciddi sakatlıkların önde gelen nedenlerinden biridir. İnme riskinin erken teşhis edilmesi, önleyici tedbirlerin alınması ve ölüm oranlarının düşürülmesi açısından hayati önem taşımaktadır. Ancak klinik verilerdeki dengesizlik, geleneksel tahmin modellerinin başarısını düşürmektedir. Bu proje, bu dengesizliği aşarak yüksek doğrulukla erken teşhis yapabilen bir yapay zeka sistemi geliştirmeyi amaçlamaktadır. Temel bilimsel soru: Mevcut klinik ölçümler, demografik bilgiler ve yaşam tarzı verileri kullanılarak; bir bireyin inme geçirme riski tahmin edilebilir mi?

1.2. Görev Türü

Bu proje, makine öğrenmesi literatüründe Gözetimli Öğrenme kategorisinde yer almaktadır. Hedef değişkenin kategorik olması ve sadece iki durum içermesi nedeniyle, problem spesifik olarak bir İkili Sınıflandırma problemidir. Veri setindeki pozitif vakaların (inme) negatif vakalara (sağlıklı) oranının çok düşük olması nedeniyle, çalışma aynı zamanda Dengesiz Veri Sınıflandırması disiplini altında ele alınmıştır.

1.3. Hedef Değişkenler

Projedeki bağımlı değişken, veri setinde stroke ismiyle yer alan sütundur. Bu değişken binary formatta olup şu sınıfları temsil etmektedir:

- 0 (Negatif Sınıf): Sağlıklı / İnme Geçirmemiş Birey.
- 1 (Pozitif Sınıf): İnme Geçirmiş / Yüksek Riskli Birey.

1.4. Başarı Kriterleri

Tıbbi tarama ve teşhis projelerinde accuracy tek başına yeterli bir başarı göstergesi değildir. Yanlış negatiflerin (hastalığın kaçırılması) maliyeti çok yüksek olduğu için projenin başarı kriterleri şu öncelik sırasına göre belirlenmiştir:

1. **Duyarlılık (Recall):** Birincil hedef, inme riski taşıyan hastaların mümkün olan en yüksek oranda tespit edilmesidir. Hedeflenen Recall oranı %80 ve üzeridir.
2. **Yorumlanabilirlik:** Modelin kararlarının tıbbi gerçeklerle örtüşmesi ve doktorlara anlamlı bir içgörü sunması gerekmektedir.

2.VERİ AÇIKLAMASI VE YÖNETİMİ

2.1. Veri Kümesi Açıklaması

Veri Seti Adı: Stroke Prediction Dataset

Kaynak: Kaggle (fedesoriano tarafından derlenmiştir).

Bağlantı: https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset

Lisans: Public / Eğitim Amaçlı Kullanım

Açıklama: Bu çalışma kapsamında kullanılan veri seti, inme risk faktörlerini analiz etmek amacıyla oluşturulmuş kapsamlı bir veritabanıdır. Veri seti, farklı yaş ve demografik özelliklere sahip bireylerden toplanan toplam 5110 adet hasta kaydından oluşmaktadır. Her bir kayıt 11 bağımsız klinik ve demografik öznitelik içermektedir. Hedef değişken, hastanın inme geçirip geçirmediğini ifade eden ikili bir yapıdadır.

2.2. Veri Şeması

Değişken Adı	Veri Tipi	Açıklama
gender	Kategorik	Cinsiyet (Erkek, Kadın, Diğer).
age	Sayısal	Hastanın yaşı.
hypertension	İkili (Binary)	Hipertansiyon geçmişi (0: Yok, 1: Var).
heart_disease	İkili (Binary)	Kalp hastalığı geçmişi (0: Yok, 1: Var).
ever_married	Kategorik	Medeni durum (Evli veya Bekar).
work_type	Kategorik	Çalışma şekli (Özel, Devlet, Serbest vb.).
Residence_type	Kategorik	İkamet türü (Kırsal veya Kentsel).
avg_glucose_level	Sayısal	Kandaki ortalama glikoz seviyesi.
bmi	Sayısal	Vücut Kitle İndeksi.
smoking_status	Kategorik	Sigara kullanım durumu.
stroke	**Hedef**	İnme durumu (0: Sağlıklı, 1: İnme).

2.3. Veri Boyutu

Toplam Satır Sayısı: 5110 Satır

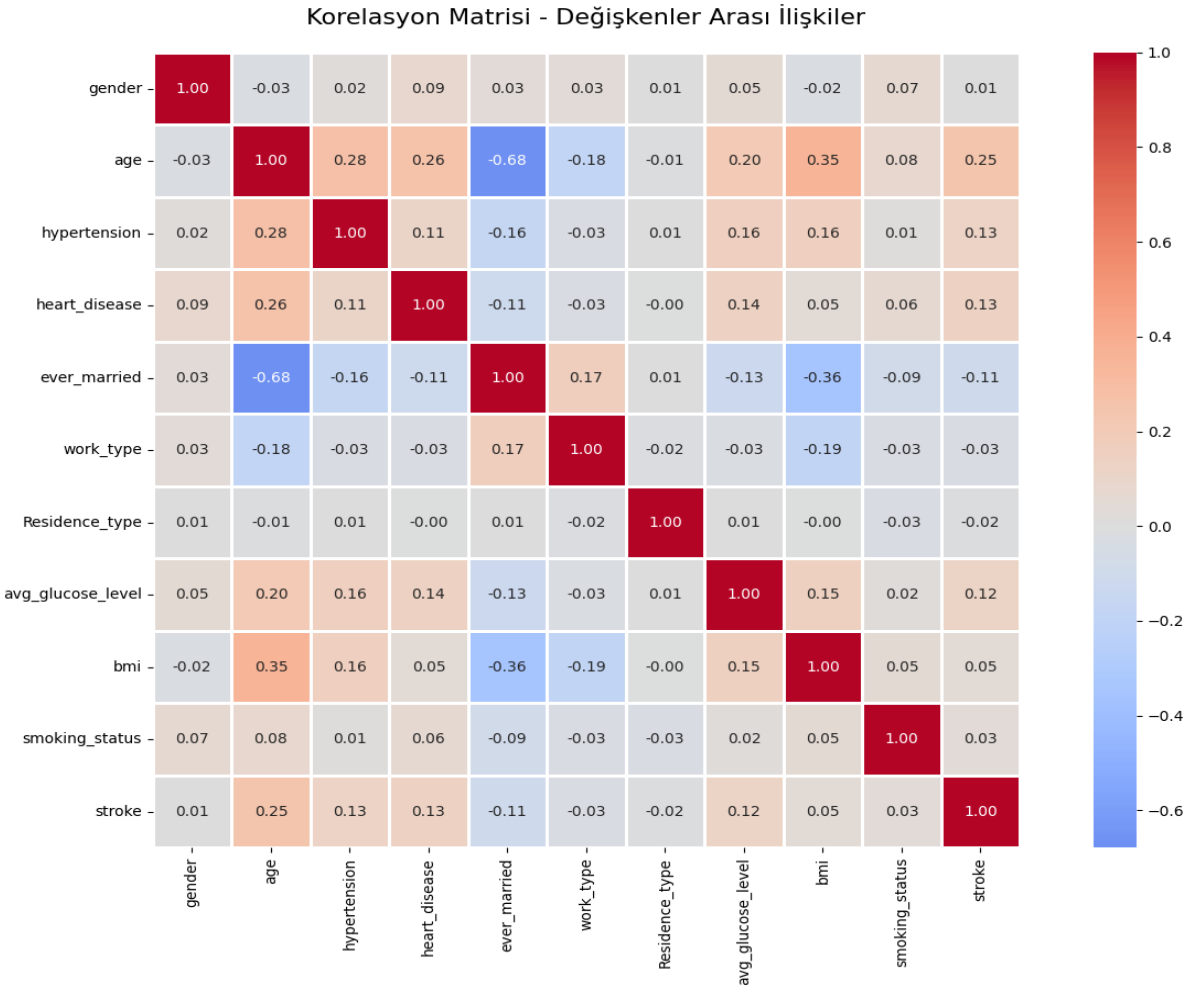
Sütun Sayısı: 12 Değişken (11 Bağımsız Değişken + 1 Hedef Değişken)

2.4. Etik ve Gizlilik

Veri seti, kimlik bilgilerinden arındırılmış anonim verilerden oluşmaktadır. id sütunu analiz öncesi çıkarılmış olup, veri setinde bireylerin gerçek kimliğini ifşa edecek herhangi bir bilgi bulunmamaktadır. Veriler yalnızca akademik araştırma ve model geliştirme amacıyla kullanılmıştır.

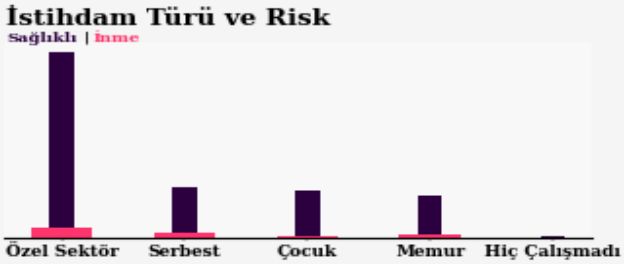
2.5. Değişkenler Arası İlişkiler

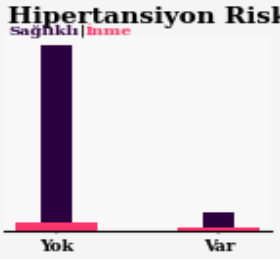
Veri setindeki bağımsız değişkenlerin birbirleriyle ve hedef değişken olan stroke ile olan doğrusal ilişkilerini incelemek amacıyla Korelasyon Katsayısı kullanılarak bir korelasyon matrisi oluşturulmuştur.

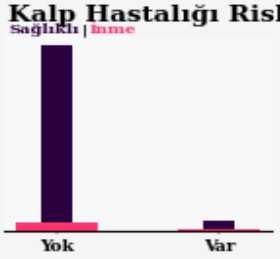


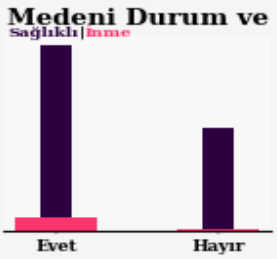
Kategorik Özelliklerin İnme Riskiyle İlişkisi

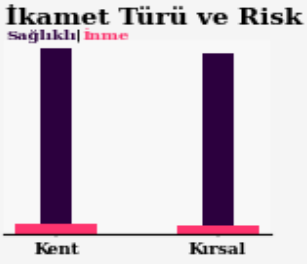
Görsel Yanılgıya Dikkat: Veri setindeki dengesizlik nedeniyle inme vakaları sayısal olarak az görünmektedir. Düşük vaka sayısı her zaman düşük riski ifade etmez. Doğru bir içgörü için mutlak sayılara değil, her grubun kendi içindeki Sağlıklı / İnme oranına odaklanılmalıdır.





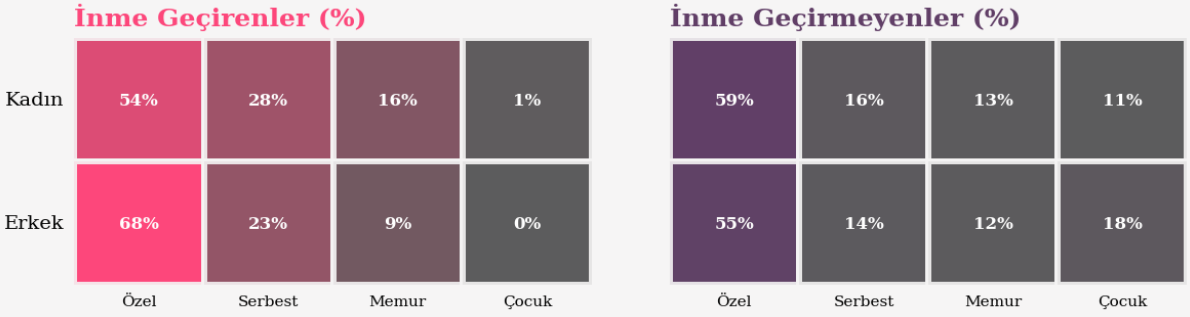






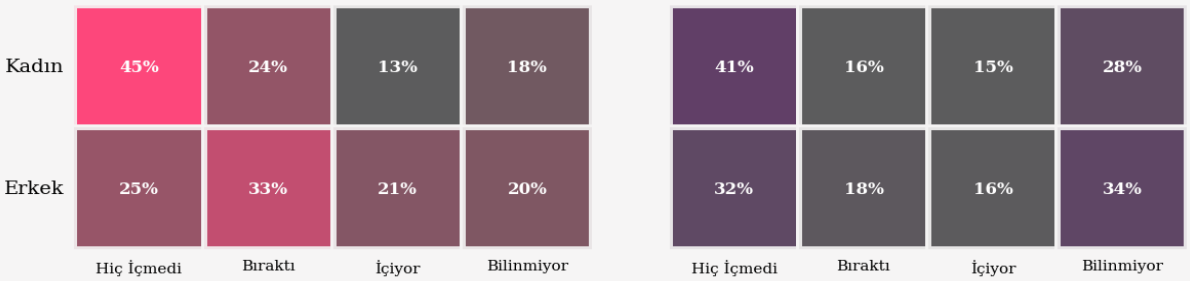
Cinsiyet ve Çalışma Şekline Göre İnme Dağılımı

Özel sektör çalışanlarının inme oranlarında en büyük paya sahip olduğu görülmektedir. Bunu serbest meslek (self-employed) grubu takip ederken, çocuklarda inme oranı beklendiği üzere oldukça düşüktür.



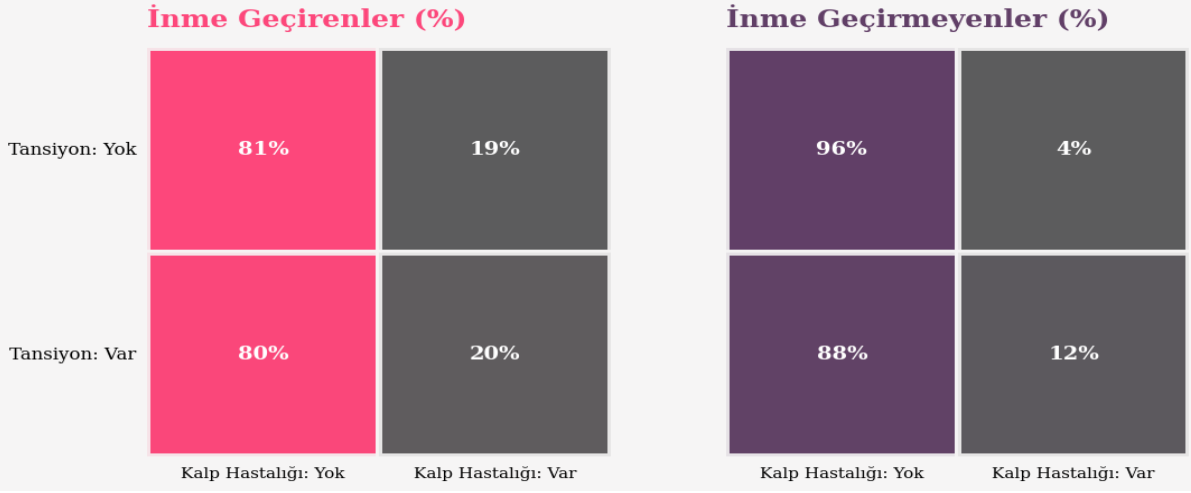
Cinsiyet ve Sigara Kullanımına Göre İnme Dağılımı

İnme vakaları incelendiğinde; kadınlarda en yüksek oran (%45) hiç sigara içmeyenlerde görülürken, erkeklerde ise sigarayı bırakmış olan grubun (%33) başı çektiği görülmektedir. Aktif içicilerin oranı ise her iki grupta da daha geridedir.



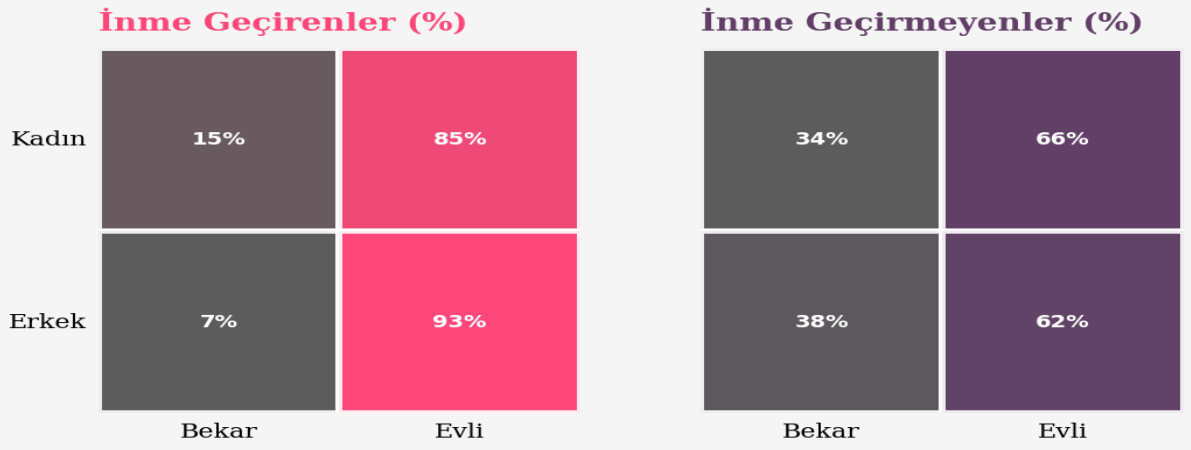
Hipertansiyon ve Kalp Hastalığına Göre Dağılım

İnme geçiren ve hipertansiyonu olan hastaların,
aynı zamanda kalp hastalığına sahip olma oranı
sağlıklı bireylere göre belirgin şekilde daha yüksektir.



Cinsiyet ve Medeni Duruma Göre İnme Dağılımı

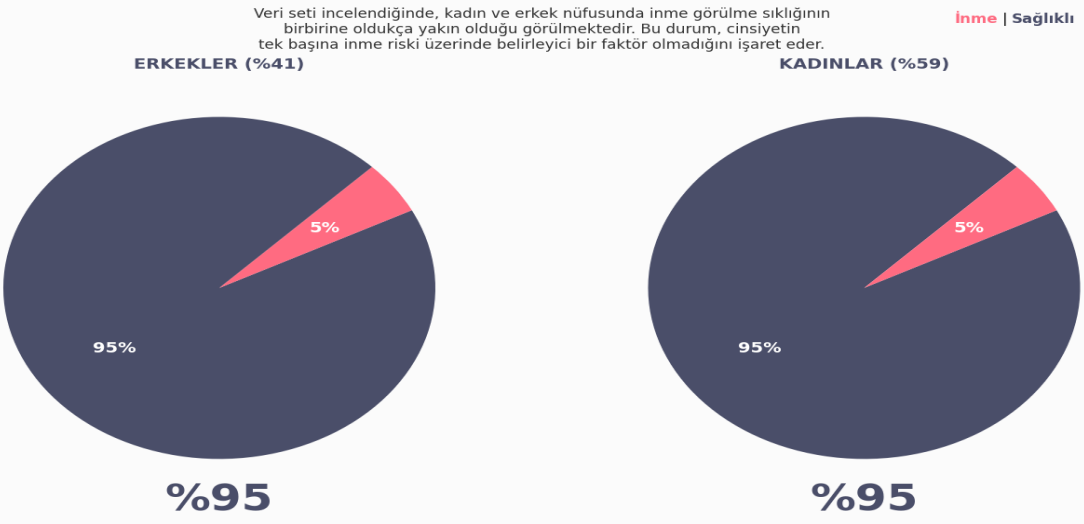
Evli bireylerin bekarlara kıyasla daha fazla inme geçirdiği görülmektedir.
Özellikle evli erkekler en çok etkilenen gruptur,
onları evli kadınlar takip etmektedir.



Cinsiyetin İnme Riski - Kadın ve Erkek Arasındaki Fark?

Veri seti incelendiğinde, kadın ve erkek nüfusunda inme görülme sıklığının
birbirine oldukça yakın olduğu görülmektedir. Bu durum, cinsiyetin
tek başına inme riski üzerinde belirleyici bir faktör olmadığını işaret eder.

İnme | Sağlıklı



3. YÖNTEMLER VE MİMARİ

3.1. Veri Ön İşleme

Eksik Veri Yönetimi: bmi değişkeninde tespit edilen eksik değerler, veri dağılımının bozulmaması adına ilgili değişkenin aritmetik ortalaması ile doldurulmuştur.

Aykırı Değer Baskılama: Sürekli değişkenlerdeki (glikoz, BMI) gürültüyü ve aşırı uç değerleri yönetmek için **Çeyrekler Arası Aralık (IQR)** yöntemi kullanılmıştır. Alt sınır ve üst sınır dışında kalan değerler, bu sınır değerlerine çekilerek baskılanmıştır.

3.2. Öznitelik Mühendisliği

Kategorik Dönüşüm: Cinsiyet, iş tipi ve medeni hal gibi nominal veriler, modelin matematiksel işlem yapabilmesi için One-Hot Encoding tekniği ile sayısal vektörlere dönüştürülmüştür.

Ölçeklendirme: Farklı birimlere sahip değişkenlerin (örn: yaş vs. glikoz seviyesi) model üzerinde baskınlık kurmasını engellemek için StandardScaler kullanılmış ve tüm veriler standart normal dağılıma çekilmiştir.

Dengesizlik Yönetimi: Eğitim verisindeki sınıf dengesizliğini gidermek için Sentetik Azınlık Aşırı Örnekleme Tekniği (SMOTE) uygulanmış ve "İnme" sınıfı yapay verilerle çoğaltılmıştır.

3.3 Model Optimizasyonu

GridSearchCV: Her bir algoritma (RF, XGB, SVM, LR) için olası hiperparametre kombinasyonları taranmış ve F1-Skoru metriğini maksimize eden en iyi parametre setleri belirlenmiştir.

Çapraz Doğrulama: Modellerin genelleme yeteneğini ölçmek için veri seti 5 parçaya bölünerek eğitim ve doğrulama işlemleri tekrarlanmıştır.

3.4. Topluluk Öğrenmesi

Mimari: Lojistik Regresyon, Random Forest, XGBoost ve SVM modellerini kapsayan bir Voting Classifier kurulmuştur.

3.5. Tahmin Katmanı

Modelin gerçek dünyada kullanılabilirliğini sağlamak amacıyla interaktif bir arayüz fonksiyonu geliştirilmiştir. Bu katman, kullanıcıdan anlık olarak alınan yaş, şeker düzeyi, BMI gibi ham verileri alır, arka planda aynı ön işleme ve ölçeklendirme adımlarından geçirir. Eğitilmiş modelleri kullanarak sonucu "Riskli" veya "Sağlıklı" olarak üretir ve güven oranını raporlar.

4. DENEY TASARIMI

4.1. Ana Amaç

Deneyin temel amacı; inme riskinin erken tahmininde hayati kritikliği göz önünde bulundurarak, "Yanlış Negatif" oranını minimize eden en güvenilir mimariyi belirlemektir.

Veri setindeki yüksek dengesizlik (%95 Sağlıklı vs %5 İnme) nedeniyle, çalışmanın odağı salt doğruluktan ziyade, azınlık sınıfı olan riskli hastaların tespitine kaydırılmıştır. Bu kapsamda; doğrusal, ağaç tabanlı, vektör tabanlı algoritmalar ve bu modellerin güçlü yönlerini birleştiren Topluluk Öğrenmesi yapısı karşılaştırmalı olarak analiz edilmiştir.

4.2. Değerlendirme Kriterleri

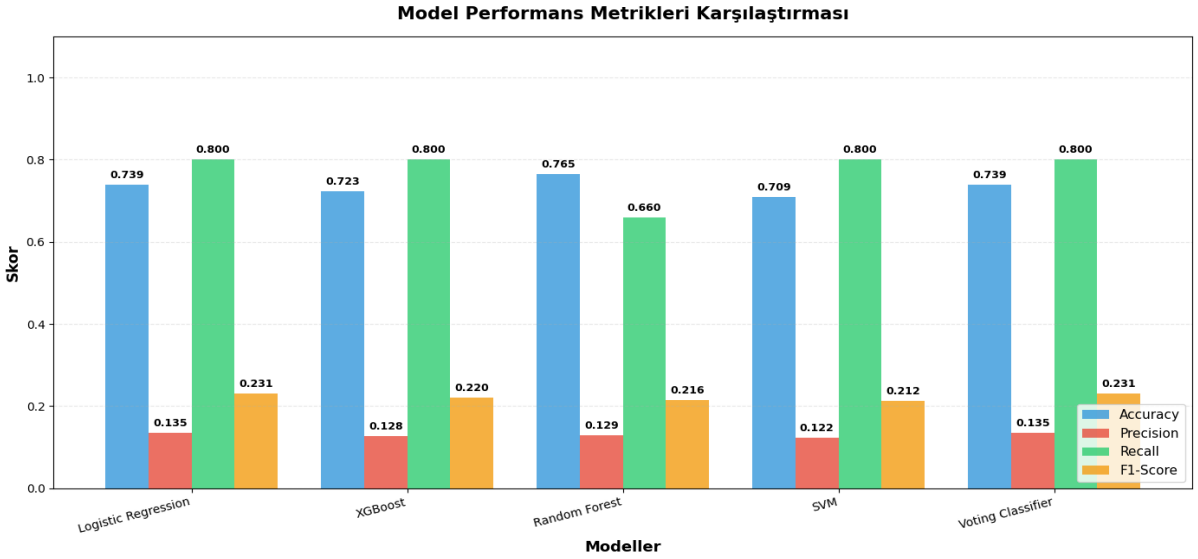
Veri setinin dengesiz yapısı nedeniyle modellerin başarısı sadece doğruluk oranıyla sınırlı kalmamış, problemin doğasına uygun şu beş temel kriterle ölçülmüştür:

- **Duyarlılık (Recall):** İnme riski taşıyan gerçek hastaların model tarafından ne oranda doğru tespit edildiği.
- **Doğruluk (Accuracy):** Sistemin genel doğru tahmin oranı.
- **Kesinlik (Precision):** Modelin "Riskli" olarak etiketlediği kişilerin gerçekte ne kadarının inme hastası olduğu.
- **F1-Skoru:** Dengesiz veri setlerinde modelin başarısını ölçmek için Precision ve Recall metriklerinin harmonik ortalaması.
- **Çapraz Doğrulama Skoru:** Modelin verinin farklı parçalarındaki tutarlılığını ölçerek, ezberleme (overfitting) riskine karşı genelleme yeteneğinin doğrulanması.

5. MODEL DEĞERLENDİRME VE BULGULAR

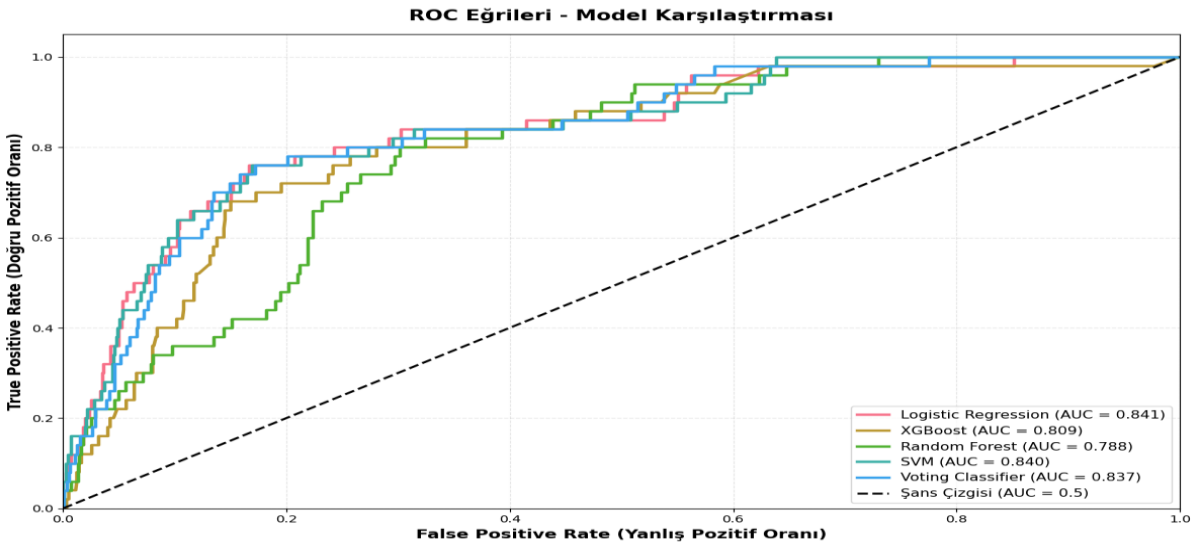
5.1. Model Performans Karşılaştırması

Modellerin performans metrikleri aşağıdaki tabloda özetlenmiştir. Tıbbi tarama testlerinin doğası gereği, değerlendirmede birincil öncelik Duyarlılık (Recall) skoruna verilmiştir.



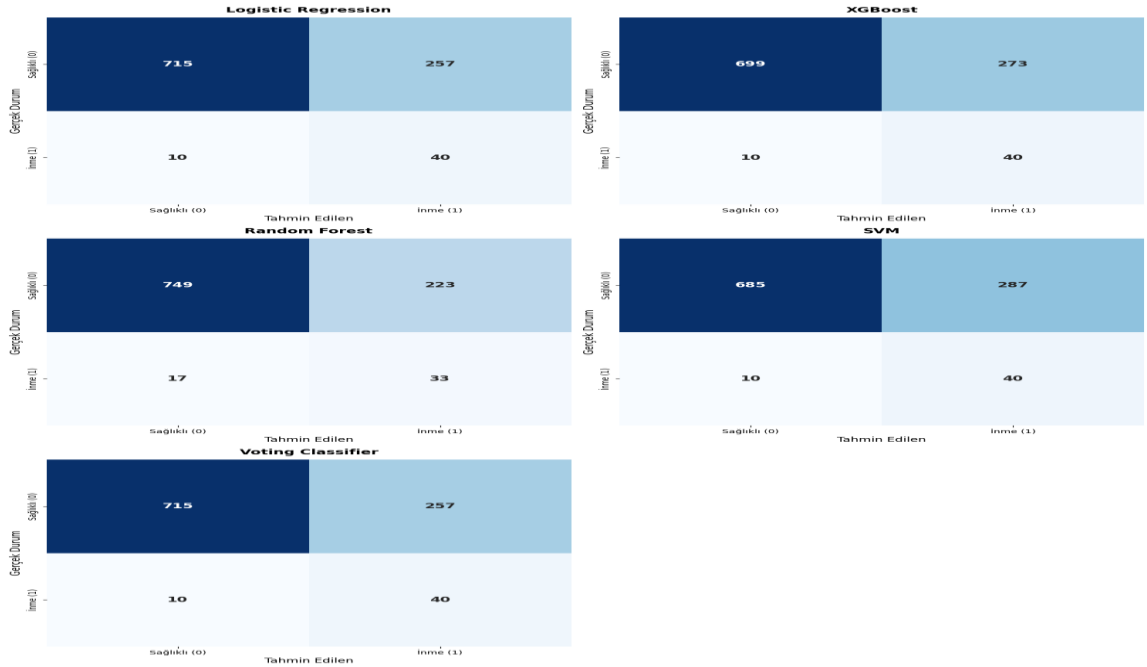
5.2. ROC Eğrisi Analizi

Grafikteki tüm modeller, şans çizgisinin üzerinde performans göstererek rastgele tahminlerden daha iyi olduklarını kanıtlamaktadır. Logistic Regression, SVM ve Voting Classifier modelleri birbirine çok yakın ve en yüksek AUC değerlerine sahipken, Random Forest modeli bu grup içinde en düşük AUC değerine sahiptir.



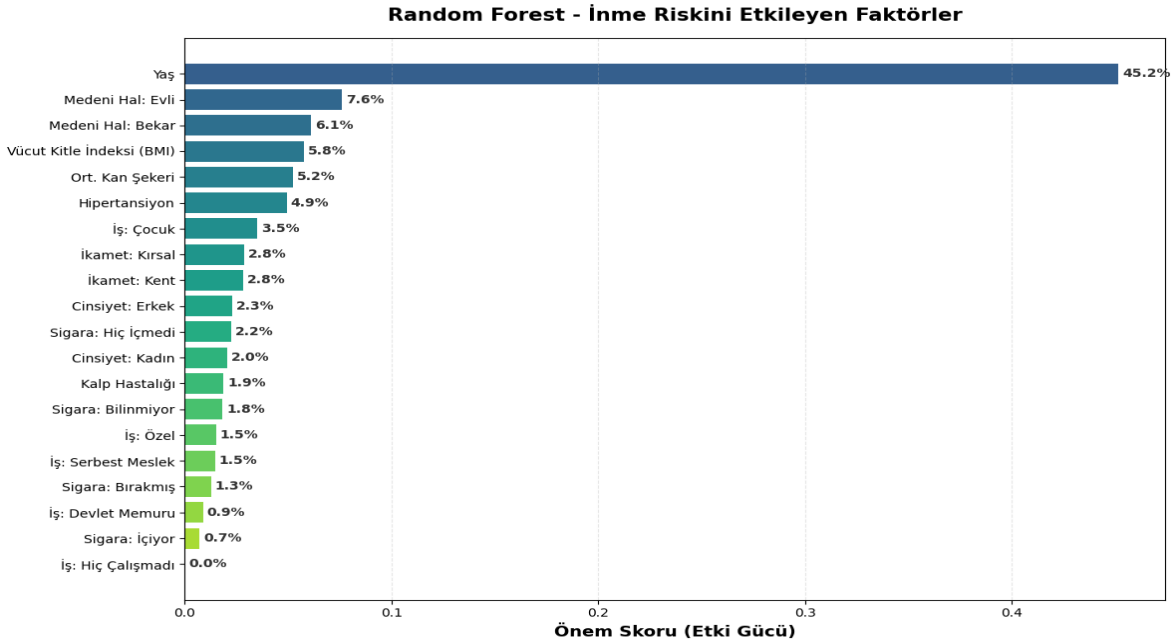
5.3. Karmaşıklık Matrisi

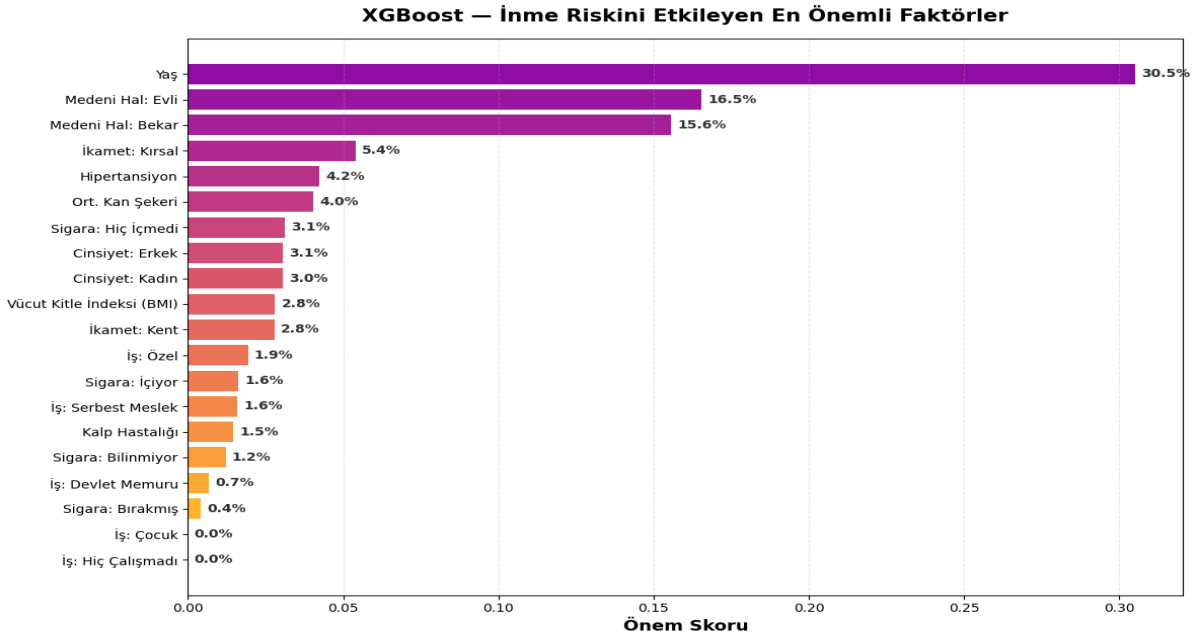
Model performansının detaylı analizi için, nihai model olarak seçilen Voting Classifier'a ait Karmaşıklık Matrisi (Confusion Matrix) incelenmiştir. Bu matris, modelin yaptığı doğru ve yanlış tahminlerin dağılımını dört temel kategoride göstermektedir.



5.4. Özellik Önem Grafiği

Modelin tahminleme sürecinde hangi klinik parametrelere öncelik verdiği, ağaç tabanlı algoritmalar üzerinden analiz edilmiştir. Bu analiz, geliştirilen yapay zekanın tıbbi literatürle ne kadar uyumlu çalıştığını kanıtlamaktadır.





5.5. Kalp Krizi Risk Tahmini

Veri setinden bağımsız, rastgele seçilen bir hasta profili oluşturulmuş ve sisteme girdi olarak verilmiştir. Bu senaryo, modelin karar destek mekanizması olarak nasıl çalıştığını simüle etmektedir.

```
================================================================================
🩺 İNME RİSKİ TAHMİN SİSTEMİ (İNTERAKTİF)
Lütfen hastanın değerlerini belirtilen aralıklarda giriniz.
================================================================================

👉 Yaş (0-120)                                                  giriniz: 72
👉 Ortalama Şeker Düzeyi (50-300)                               giriniz: 230
👉 Vücut Kitle İndeksi (BMI) (10-60)                            giriniz: 38.5
👉 Hipertansiyon (1: Var, 0: Yok)                               giriniz: 1
👉 Kalp Hastalığı (1: Var, 0: Yok)                              giriniz: 1
👉 Cinsiyet (1: Erkek, 0: Kadın)                                giriniz: 1
👉 Evlilik Durumu (0: Evli, 1: Bekar)                           giriniz: 0
👉 İş (0:Özel, 1:Serbest, 2:Çocuk, 3:Memur, 4:Çalışmadı)        giriniz: 0
👉 İkamet (0: Kent, 1: Kırsal)                                  giriniz: 0
👉 Sigara (0:Hiç, 1:Bilinmiyor, 2:Bırakmış, 3:İçiyor)           giriniz: 3

================================================================================
🤖 MODEL TAHMİN SONUÇLARI
================================================================================
Model Adı                 | Tahmin                  | Güven Oranı
------------------------------------------------------------------
Logistic Regression       | İNME RİSKİ VAR 🚨       | %95.69 Oran.
XGBoost                   | İNME RİSKİ VAR 🚨       | %68.18 Oran.
Random Forest             | İNME RİSKİ VAR 🚨       | %80.32 Oran.
SVM                       | İNME RİSKİ VAR 🚨       | %96.39 Oran.
Voting Classifier         | İNME RİSKİ VAR 🚨       | %85.15 Oran.
```

6. GELİŞTİRME ORTAMI

- **Python 3.x:** Projenin geliştirildiği temel programlama dili ve çalışma ortamı.
- **Pandas & NumPy:** Veri setinin yüklenmesi, eksik verilerin yönetimi, matris hesaplamaları ve vektörel işlemler gibi süreçlerde kullanılmıştır.
- **Scikit-Learn (Sklearn):** Makine öğrenmesi mimarisinin ana omurgasını oluşturur. Veri ölçeklendirme, eğitim-test ayrımı, model eğitimi , GridSearchCV ile hiperparametre optimizasyonu ve Topluluk Öğrenmesi yapısı bu kütüphane ile kurgulanmıştır.
- **Imbalanced-Learn:** Veri setindeki sınıf dengesizliğini gidermek amacıyla kullanılan SMOTE algoritması ve veri sızıntısını önleyen özel Pipeline yapısı bu kütüphane aracılığıyla projeye entegre edilmiştir.
- **XGBoost:** Yüksek hız ve performans sunan, gradyan artırma tabanlı sınıflandırma algoritması olarak kullanılmıştır.
- **Matplotlib, Seaborn & Plotly:** Keşifsel veri analizi sırasında veri dağılımlarının incelenmesi; sonuç bölümünde yer alan Karmaşıklık Matrisi, ROC Eğrisi ve Özellik Önemi grafiklerinin oluşturulması için bu kütüphanelerden yararlanılmıştır.

7. SONUÇ VE DEĞERLENDİRME

Projeye verilerimizi keşfederek başladık. İlk etapta yaş, ortalama glikoz düzeyi ve VKİ gibi sayısal verilerin güçlü birer gösterge olduğunu; aynı zamanda medeni hal, çalışma şekli ve sigara kullanımı gibi kategorik faktörlerin de inme riskiyle anlamlı ilişkiler taşıdığını tespit ettik.

Kapsamlı veri görselleştirme süreçlerinin ardından, veri setindeki dengesizliği gidermek için SMOTE tekniği uygulandı ve problemi çözmek için birden fazla makine öğrenmesi algoritması test edildi. Bu süreçte Random Forest, SVM, XGBoost ve Lojistik Regresyon algoritmaları eğitilerek performansları karşılaştırıldı. Modellerin ham sonuçlarını aldıktan sonra, başarı oranlarını daha da artırmak amacıyla Hiperparametre Optimizasyonu uyguladık.

Karşılaştırma sonucunda; Random Forest modeli %76.52 ile en yüksek doğruluk (accuracy) oranına sahip olsa da, Voting Classifier modeli %80.00 Duyarlılık (Recall) ve %23.05 F1 Skoru ile inme vakalarını tespit etmede en başarılı sonucu verdi. İnme teşhisinde 'riskli hastayı gözden kaçırmamak' projenin ana hedefi olduğu için, nihai model olarak Voting Classifier mimarisini tercih ettik. Ancak analiz süreci bununla sınırlı kalmadı. Modelin verileri nasıl kullandığını ve kararlarını hangi mantığa dayandırdığını anlamak için Özellik Önem Düzeylerini inceleyerek çalışmayı tamamladık.

8. KAYNAKLAR

[1] Veri Seti Kaynağı: Fedesoriano. "Stroke Prediction Dataset." Kaggle. *Erişim Adresi:* https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset

[2] Tanmay Deshpande. "Stroke Prediction: Effect of Data Leakage & SMOTE." Kaggle Kernels. *Erişim Adresi:* https://www.kaggle.com/code/tanmay111999/stroke-prediction-effect-of-data-leakage-smote

[3] Joshua Swords. "Predicting a Stroke: SHAP, LIME, Explainer, ELI5." Kaggle Kernels. *Erişim Adresi:* https://www.kaggle.com/code/joshuaswords/predicting-a-stroke-shap-lime-explainer-eli5

[4] Bhuvan Chennoju. "Data Storytelling ~ AUC Focus on Strokes." Kaggle Kernels. *Erişim Adresi:* https://www.kaggle.com/code/bhuvanchennoju/data-storytelling-auc-focus-on-strokes

[5] Nur Ahmadi. "Stroke-prediction-with-ML." GitHub Repository. *Erişim Adresi:* https://github.com/nurahmadi/Stroke-prediction-with-ML