Sağlık Verileri ile İnme Riskinin Tahmini: Veri Analizi ve Makine Öğrenimi Uygulamaları

**Giriş**

Bu proje, inme (stroke) riskini tahmin etmek amacıyla sağlık verileri kullanılarak bir makine öğrenimi modeli geliştirmeyi hedeflemektedir. İnme, dünya genelinde önemli bir sağlık sorunu olup, erken teşhis ve önleyici tedbirlerin alınması hayati öneme sahiptir. Proje kapsamında, veri analizi ve modelleme adımları izlenerek doğruluğu yüksek bir tahmin modeli oluşturulmuştur.

**Kullanılan Yöntemler**

Projede, veri ön işleme, veri dengeleme, modelleme ve değerlendirme adımları izlenmiştir:

1. **Veri Ön İşleme:** Eksik değerlerin doldurulması, kategorik değişkenlerin sayısal formata dönüştürülmesi gibi işlemler yapılmıştır.
2. **Veri Dengeleme:** SMOTE (Synthetic Minority Over-sampling Technique) kullanılarak veri setindeki sınıf dengesizliği giderilmiştir.
3. **Özellik Ölçekleme:** Özellikler StandardScaler kullanılarak standartlaştırılmıştır.
4. **Modelleme:** Logistic Regression kullanılarak model eğitimi gerçekleştirilmiştir.
5. **Değerlendirme:** Modelin performansı doğruluk, hassasiyet, duyarlılık ve F1 skoru ile değerlendirilmiştir.

**Veri Seti Hakkında Bilgi**

Veri seti, inme geçiren ve geçirmeyen bireylerin çeşitli sağlık ve demografik bilgilerini içermektedir. Kullanılan veri setinde aşağıdaki değişkenler bulunmaktadır:

* id: Bireysel kimlik numarası
* gender: Cinsiyet (Male, Female, Other)
* age: Yaş
* hypertension: Yüksek tansiyon durumu
* heart\_disease: Kalp hastalığı durumu
* ever\_married: Evlilik durumu
* work\_type: Çalışma durumu (Private, Self-employed, Govt\_job, Children, Never\_worked)
* Residence\_type: Yaşam yeri (Urban, Rural)
* avg\_glucose\_level: Ortalama glukoz seviyesi
* bmi: Vücut kitle indeksi
* smoking\_status: Sigara içme durumu (formerly smoked, never smoked, smokes, unknown)
* stroke: İnme durumu (Evet: 1, Hayır: 0)

Eksik değerler sadece bmi değişkeninde bulunmaktadır ve bunlar sütunun medyanı ile doldurulmuştur. Kategorik değişkenler one-hot encoding yöntemi ile sayısal formata dönüştürülmüştür.

**Veri Dağılımları**

Aşağıdaki görseller, veri setindeki sayısal ve kategorik değişkenlerin dağılımlarını göstermektedir:

**Sayısal Değişkenlerin Dağılımı:** A screenshot of a graph

Description automatically generated

**Kategorik Değişkenlerin Dağılımı:**

A group of colorful squares

Description automatically generated with medium confidence

**Deneyde Kullanılan Modeller / Mimariler**

Bu projede Logistic Regression kullanılmıştır. Logistic Regression, sınıflandırma problemlerinde yaygın olarak kullanılan ve basit fakat etkili bir modeldir. Model, eğitim ve test veri setlerine ayrılarak doğruluğu artırılmış ve performansı değerlendirilmiştir.

* **Model:** Logistic Regression
* **Özellik Ölçekleme:** StandardScaler
* **Veri Bölme:** Training (%60), validation (%20) ve test (%20) olarak ayrılmıştır.
* **Özellik Seti:** İnme durumu dışındaki tüm sütunlar özellik olarak kullanılmıştır.

**Deney Sonuçları**

Modelin performansı aşağıdaki metriklerle değerlendirilmiştir:

* **Doğruluk (Accuracy):** 0.9568
* **Hassasiyet (Precision):** 0.9944
* **Duyarlılık (Recall):** 0.9186
* **F1 Skoru (F1 Score):** 0.9550

**Karışıklık Matrisi**

Modelin karışıklık matrisi, sınıflandırma başarısını daha detaylı bir şekilde göstermektedir:

A blue squares with white text

Description automatically generated

**Tartışma**

Bu çalışma, Logistic Regression modelinin veri seti üzerinde oldukça başarılı bir şekilde çalıştığını göstermektedir. Model, inme tahmininde yüksek bir doğruluk ve F1 skoru ile dikkat çekmektedir.

**Eksiklikler**

1. **Karmaşık Modellerin Eksikliği:** Daha karmaşık modellerin kullanılması, tahmin performansını daha da artırabilir.
2. **Veri Seti Kapsamı:** Veri seti, yalnızca belirli özellikleri içermekte olup, daha fazla özellik eklenmesi performansı artırabilir.
3. **Eksik Veri Sorunları:** Eksik veri doldurma işlemi model performansını etkileyebilir; daha kaliteli veri seti ile daha iyi sonuçlar elde edilebilir.

**Güçlü Yönler**

1. **Veri Dengeleme:** SMOTE ile sınıf dengesizliği başarılı bir şekilde giderilmiştir.
2. **Özellik Ölçekleme:** Özelliklerin ölçeklendirilmesi modelin performansını artırmıştır.
3. **Yüksek Performans:** Model, yüksek doğruluk ve F1 skoru ile iyi bir sınıflandırma başarısı göstermiştir.

Gelecekteki çalışmalar, daha karmaşık modellerin denenmesi ve veri setine yeni özelliklerin eklenmesiyle daha yüksek doğruluk elde edebilir.

**Referanslar**

1. “Synthetic Minority Over-sampling Technique.” Wikipedia, Wikimedia Foundation, <https://en.wikipedia.org/wiki/Synthetic_minority_over-sampling_technique>.
2. “Logistic Regression.” Wikipedia, Wikimedia Foundation, <https://en.wikipedia.org/wiki/Logistic_regression>.