Veri Bilimi Vize Ödevi Raporu

Feyzullah Durası 221041018

Konu: Sigorta Ücreti Tahmini Yapmak

1. Kullanılan Veri Seti Hakkında Bilgi

Bu projede kullanılan veri seti, bireylerin demografik bilgileri (yaş, cinsiyet, vücut kitle indeksi vb.) ve sağlık durumlarına ilişkin bilgileri içermektedir.
Veri setinin amacı, bireylerin sigorta ücretlerini tahmin etmektir. Veri seti aşağıdaki değişkenleri içermektedir:

- age: Kişinin yaşı, sex: Cinsiyet
- bmi: Vücut kitle indeksi
- children: Sahip olunan çocuk sayısı
- smoker: Sigara kullanımı
- region: Bölge bilgisi
- charges: Sigorta ücreti (hedef değişkenimiz)

Veri seti yüklendikten sonra keşifsel veri analizi işlemleri yapılmıştır. Verinin genel istatistikleri incelenmiş, eksik veri bulunmadığı tespit edilmiştir.
Ayrıca, hedef değişken olarak charges yani sigorta ücreti belirlenmektedir.

2. Veri Ön işleme

Veri ön işleme adımları şunlardır:

- Eksik Verilerin İşlenmesi: Veri setinde eksik veri bulunmadığı için herhangi bir doldurma ya da çıkarma işlemi yapmadım.

- Kategorik Verilerin Sayısallaştırılması:
  - sex, smoker ve region değişkenleri kategorik olduğundan Label Encoding yöntemi ile sayısal hale getirilmiştir.

- Veri Normalizasyonu:
  - Özellikle yaş, bmi ve charges gibi sayısal değişkenlerin daha iyi model performansı sağlaması için StandardScaler ile standardizasyon işlemi uygulanmıştır.

3. Model Seçimi ve Eğitimi

Veri setinin yapısı ve hedef değişkenin sürekli (sayısal) bir değişken olması nedeniyle regresyon problemlerine uygun modeller tercih edilmiştir.
Bu doğrultuda birkaç farklı model denenmiş ve performansları karşılaştırılmıştır:

- Linear Regression
- Random Forest Regressor
- Lasso Regression
- Ridge Regression

Modellerin eğitimi sırasında veri eğitim ve test olarak ayrılmıştır. train_test_split işlemiyle %80 eğitim, %20 test.

4. Model Değerlendirme

Model başarılarını değerlendirmek için aşağıdaki metrikler kullanılmıştır:

- R^2 (Determination Coefficient)
- Mean Absolute Error (MAE)
- Mean Squared Error (MSE)

Model sonuçlarına göre, Random Forest Regressor modeli sigorta ücreti tahmininde en başarılı model olmuştur. Bu model, diğer lineer modellere göre daha düşük ortalama hata değerlerine ve daha yüksek doğruluk oranına sahiptir.

- Karmaşık ilişkilere uyum sağlayabilen Random Forest modeli, verideki doğrusal olmayan yapıları daha iyi öğrenmiştir.
- Aşırı öğrenme eğilimi azdır ve genel performansı yüksektir.

Model	R^2 Skoru	MAE	MSE
Linear Regression	0.76	4186.08	35814265.0
Ridge Regression	0.76	4186.06	35814518.0
Lasso Regression	0.76	4185.95	35814233.0
Random Forest R.	0.86	2533.72	21761729.0

5. Notlar (Ödevi yaparken öğrendiğim kazanımlar)

- **R^2 (Determination Coefficient)**: Modelin verideki toplam değişkenliğin ne kadarını açıkladığını gösteren doğruluk ölçüsüdür; 1'e ne kadar yakınsa model o kadar başarılıdır.
- **Mean Absolute Error (MAE)**: Tahmin edilen değerler ile gerçek değerler arasındaki mutlak farkların ortalamasıdır; hata miktarını sade bir şekilde gösterir.
- **Mean Squared Error (MSE)**: Hataların karesinin ortalamasıdır; büyük hataları daha fazla cezalandırır.
- **Lasso Regression**: Özellik seçiminde etkili olan, bazı katsayıları sıfıra indirerek daha sade modeller oluşturan bir regresyon yöntemidir.
- **Ridge Regression**: Aşırı öğrenmeyi (overfitting) engellemek için model katsayılarını küçülterek genelleme yeteneğini artıran bir regresyon tekniğidir.