Maliyet Analizi Projesi*

*Not: Proje Global AI Hub AI Summer Camp'22 etkinliği kapsamında Pytorturers proje grubu tarafından hazırlanmıştır.

Hakan Özdemir
İstanbul, Türkiye
ozdemir.hkn@outlook.com

Baha Özşahin
Bursa, Türkiye
bahaozsahin@gmail.com

Esra Nur Erkek
Isparta, Türkiye
esranurerkek32@gmail.com

Melisa Çağılgan
Tekirdağ, Türkiye
melisacagilgan@gmail.com

Abdurrahman Keskin
İstanbul, Türkiye
abdurrahman.ksn@gmail.com

I. GİRİŞ

Bu projede, Sağlık Sigortası Maliyeti Veri Seti [1] kullanılarak uçtan uca bir veri bilimi uygulaması geliştirmeye çalışılmıştır. Projenin amacı, verilen değişkenlere göre bir kişinin sağlık sigortasının yaklaşık ne kadar masraflı olacağını tahmin etmektir. Bu amacı gerçekleştirmek için, veri kümesine uygulanabilecek çeşitli Makine Öğrenimi teknikleri ve veri kümesinin analizi bu çalışmada incelenmektedir.

II. VERİ SETİ

Bu projede, Kaggle [2] veri platformunda bulunan Sağlık Sigortası Maliyeti Veri Seti [1] kullanılmıştır. Veri setine ait değişken ve açıklamaları Tablo 1 'de verilmiştir.

Tablo 1: Veri seti değişkenleri ve açıklamaları

Değişken	Açıklama
age	yaş
sex	cinsiyet
bmi	vücut kitle indeksi
children	sahip olduğu çocuk sayısı
smoker	sigara içip içmediği
region	bulunduğu bölge
charges	ödenen ücret bilgisini içeriyor.

III. VERİYİ ANALİZ ETME

Veriler incelenip, analiz ederek veriden anlamlı sonuçlar çıkartılmıştır.

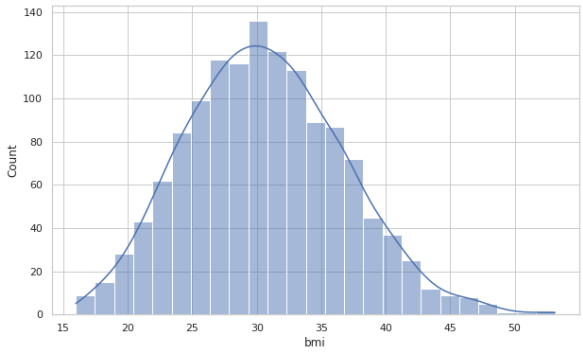
```
############################################################
İlk 5 veri
   age     sex     bmi  children smoker     region      charges
0   19  female  27.900         0    yes  southwest  16884.92400
1   18    male  33.770         1     no  southeast   1725.55230
2   28    male  33.000         3     no  southeast   4449.46200
3   33    male  22.705         0     no  northwest  21984.47061
4   32    male  28.880         0     no  northwest   3866.85520
############################################################
Verinin boyutları:
(1338, 7)
```

```
############################################################
Verideki boş gözlem sayısı:
age         0
sex         0
bmi         0
children    0
smoker      0
region      0
charges     0
dtype: int64
############################################################
Verideki kategorik değişkenler
['sex', 'smoker', 'region', 'children']
############################################################
Verideki sayısal değişkenler
['age', 'bmi', 'charges']
```

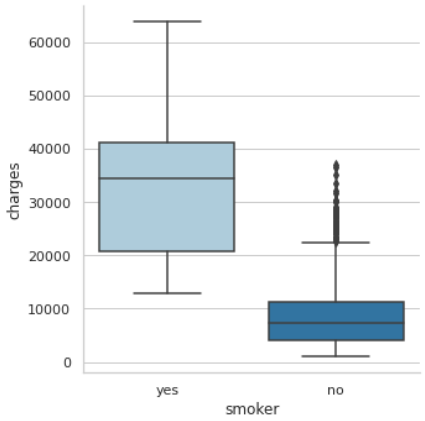
Şekil 1. Veri setine genel bir bakış

IV. VERİYİ GÖRSELLEŞTİRME

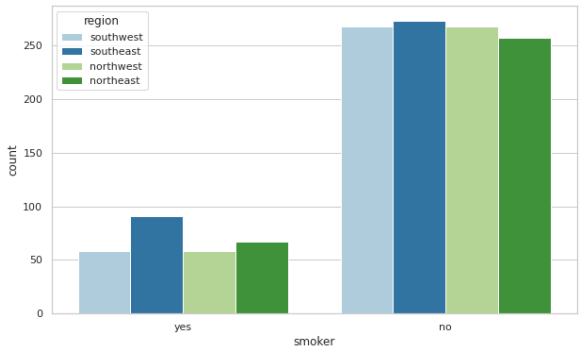
- Bmi(Vücut Kitle İndeksi)'nin dağılımını:



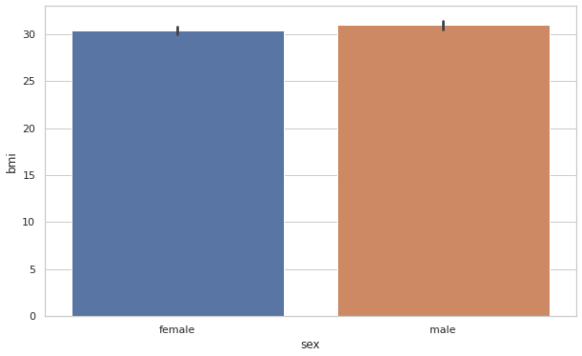
- "smoker" ile "charges" arasındaki ilişki:

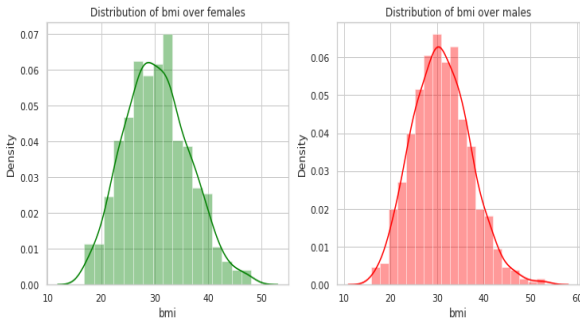


- “smoker” (Sigara tüketen) ile “region”(Bölge) arasındaki ilişki:

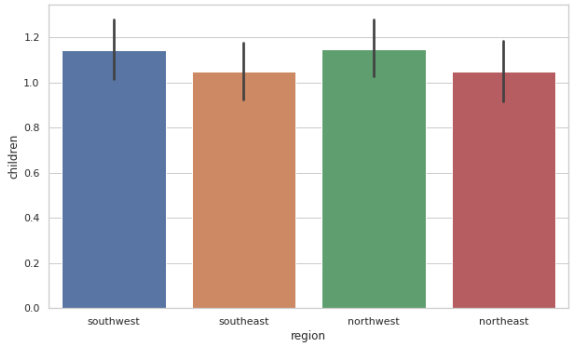


- “bmi” ile “sex”(Cinsiyet) arasındaki ilişki:





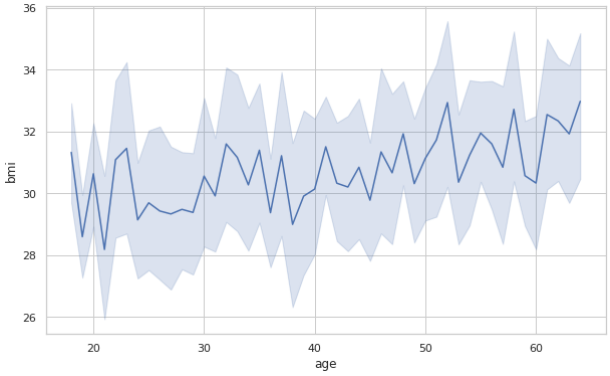
- En çok “children”’a sahip “region”:



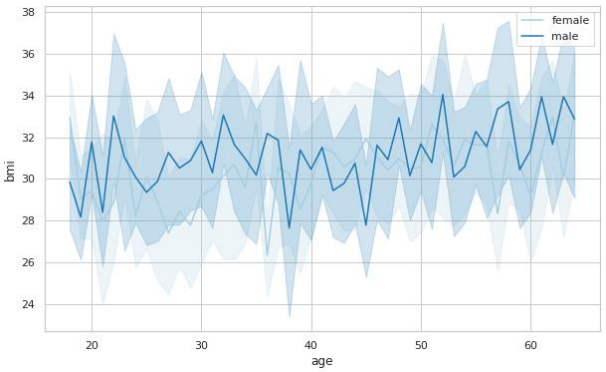
region	count
northeast	324
northwest	325
southeast	364
southwest	325

region	count
southeast	364

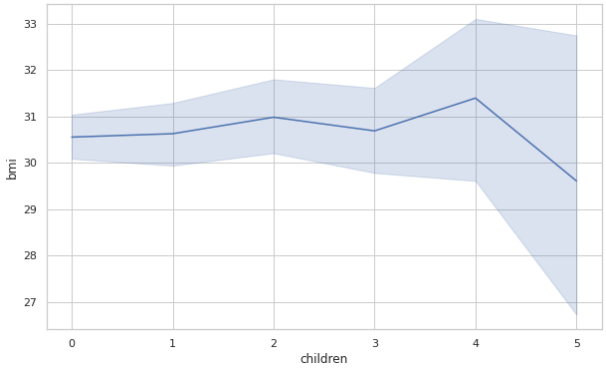
- “Age” ile “bmi” arasındaki ilişki:



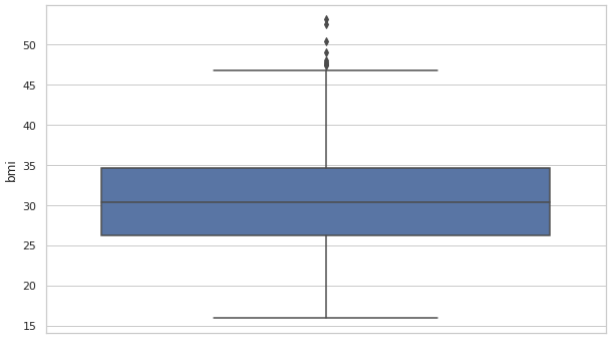
“sex” “age” ve “bmi” bazında da arasında ilişki:



- “bmi” ile “children” arasındaki ilişki:

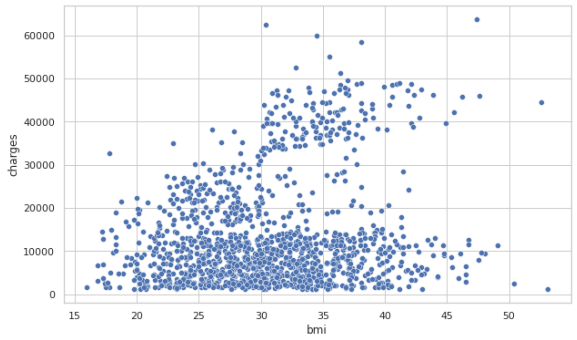


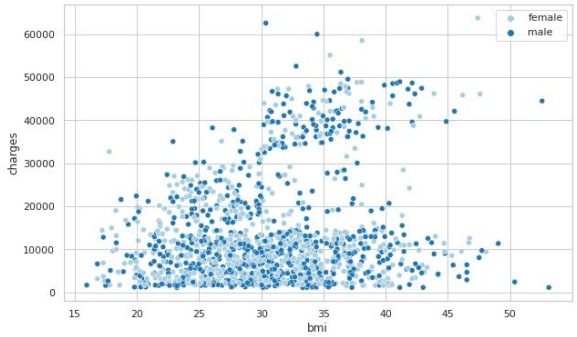
- “bmi” değişkeninde outlier var mıdır?



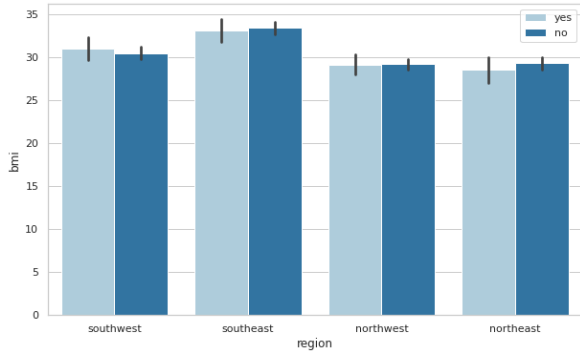
Yukarıdaki grafikte de görüldüğü üzere “bmi” değişkeninde bazı aykırı değerler olduğu gözlemlenmiştir.

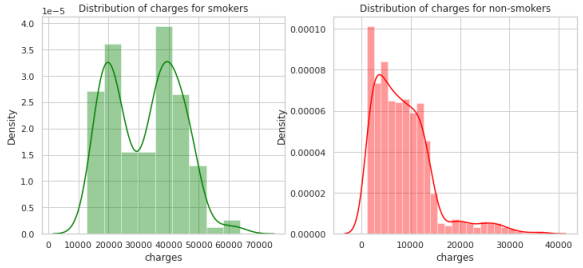
- “bmi” ile “charges” arasındaki ilişki:





- “region”, “smoker” ve “bmi” arasındaki ilişki:





V. VERİ ÖN İŞLEME

Makine Öğrenimi modeli, ondan kalıpları çıkarmak ve tahminler yapmak için sayısal gösterimlerde girdi verileri gerektirir. Ancak kaynak veri kümemizde sağlanan tüm veriler sayısal değildir. Sağlanan verilerden bazıları sex, smoker, region kategorik verilerdir. Bunları sayısal gösterimlere dönüştürmemiz gerekir.

Burada veriler, modelimizin girdi olarak kullandığı bir özellikten başka bir şey değildir. Bu nedenle, kaynak veri kümesinden anlamlı sayısal veriler oluşturmak için verilerimiz üzerinde güncellemeler yapılmalıdır.

Elimizde 7 değişken bulunmaktadır. Nihai hedefimiz, “sigorta için ödenen ücret bilgisi” sorusunun cevabını bağımlı değişkenimiz yapan charges özelliğini tahmin etmektir. Kalan 6 değişkenden herhangi biri hedef değişkeni etkiliyorsa, bu özellikler bağımsız değişkenler olarak bilinir.

Kategorik veriler ML (Machine Learning) modelinin anlayabileceği sayısal gösterimlere one hot encoding ve label encoding teknikleri kullanılarak dönüştürüldü.

Region değişkeni için one hot encoding yöntemi, sex, smoker değişkenleri için label encoding yöntemi kullanılmıştır.

- Label encoding yöntemi kullanılarak dönüştürülmüş sex değişkeni:

sex	
female	**male**
0	1

- Label encoding yöntemi kullanılarak dönüştürülmüş smoker değişkeni:

smoker	
no	**yes**
0	1

- One hot encoding yöntemi kullanılarak dönüştürülmüş region değişkeni:

region			
northeast	**northwest**	**southeast**	**southwest**
1000	0100	0010	0001

VI. VERİ NORMALLEŞTİRMESİ

Normalleştirme, genellikle makine öğrenimi için veri hazırlamanın bir parçası olarak uygulanmaktadır. Normalleştirmenin amacı, veri kümesindeki sayısal sütunların değerlerini, değer aralıklarındaki farklılıkları bozmadan ortak bir ölçeğe değiştirmektir.

Değişkenlerde 0 ve 1 arasındaki değer ataması yapılmıştır.

$$x' = (x - x_{min}) / (x_{max} - x_{min})$$

	age	sex	bmi	children	smoker	charges	region_northeast	region_northwest	region_southeast	region_southwest
0	0.021739	0	0.321227	0.0	1	0.251611	0	0	0	1
1	0.000000	1	0.479150	0.2	0	0.009636	0	0	1	0
2	0.217391	1	0.458434	0.6	0	0.053115	0	0	1	0
3	0.326087	1	0.181464	0.0	0	0.333010	0	1	0	0
4	0.304348	1	0.347592	0.0	0	0.043816	0	1	0	0

VII. VERİLERİ BÖLME

Makine Öğrenimi modeli, ondan kalıpları çıkarmak ve tahminler yapmak için sayısal gösterimlerde girdi verileri gerektirir. Ancak kaynak veri kümemizde sağlanan tüm veriler sayısal değildir. Sağlanan verilerden bazıları sex,

smoker, region kategorik verilerdir. Bunları sayısal gösterimlere dönüştürmemiz gerekir.

Veri setini 8,5-1,5, 8-2, ve 7,5-2,5 şeklinde eğitim ve test verileri olarak bölerek tüm sonuçları karşılaştırdık. En yüksek doğruluk payı değerine eğitimi 8,5 (%85) ve test verilerini 1,5 (%15) şeklinde böldüğümüzde ulaştığımızı fark ettiğimiz için bölme işlemini bu doğrultuda gerçekleştirdik.

VIII. MODEL SEÇİMİ

Regresyon, kısaca bir bağımlı değişken ve bir veya birden fazla bağımsız değişkenin ilişkisini istatistiksel olarak incelemeye yarayan bir metottur. Araştırmalarımız sonucunda üstünde çalıştığımız problemin ve buna benzer sorunların regresyon yöntemleri ile çözüldüğünü gördük. Regresyon modeli seçimi sırasında 7 farklı model eğittik. Eğitilen modellerden optimale en yakın sonucu veren modelin seçilmesini uygun bulduk. Eğitilen modellerden bazıları, Lineer, Ridge, Bayesian Ridge ve Destek Vektör Regresyonu yöntemleri olarak sayılabilir.

IX. HİPER-PARAMETRE OPTİMİZASYONU

Proje kapmasında bizden birçok regresyon modeli denememiz istendiği için bir önceki adımda hangi modelin seçileceğini bilmediğimizi varsayarak kullandığımız tüm modeller ile çalışacak şekilde bir hiper-parametre optimizasyonu hazırladık. Bu şekilde bir optimizasyon hazırlamak için öncelikle tüm modellerin kullandığı parametreleri tek tek inceledik. Daha sonra uygun gördüğümüz parametrelere çeşitli değişkenler ekledik ve optimizasyonumuzu tüm modellerimiz için kullanılabilir hale getirdik ve en iyi model üzerinde kullandık. Buna ek olarak tüm modeller üzerinde çalışıp çalışmadığını da kısa bir for döngüsü kullanarak test ederek projemizin sorunsuz çalıştığından emin olduk.

X. MODELİ DEĞERLENDİRME

Modeli değerlendirirken çapraz doğrulama yöntemi araştırmalarımız doğrultusunda bulduğumuz regresyon için kullanılan 3 adet metriği kullandık. Önce çapraz doğrulama sonuçlarını kullanarak en iyi modeli bulduk. Daha sonra kendi belirlediğimiz metrikleri kullanarak bir model belirledik ve iki sonucu karşılaştırarak aynı olup olmadıklarına baktık. Belirlediğimiz metriklerden 2'si hata payını gösteren Mean Squared Error (MSE) ve Mean Absolute Error (MAE) iken 1'i ise skor göster R^2'di. Modeli değerlendirirken öncelikle tüm metriklerin sadece hata ya da sadece doğruluk payı göstermesine dikkat ettik ve R^2 değerini 1'den çıkararak onu da hata payına çevirdik ve 3 metriğin ortalamasını alarak ortalama hata payını veren bir metrik elde ettik. İki değerlendirmenin sonuçlarını karşılaştırdık ve ikisinin de Histogram-Based Gradient Boosting Regression Tree modelini seçtiklerini gördük. Bu yüzden biz de çalışmamızı bu modeli kullanarak yaptık.

XI. SONUÇ

Veri setimizin ve seçtiğimiz en iyi modelin üzerinde uyguladığımız bütün teknikler sonucu doğruluk payı %88,50 olarak bulunmuştur. Modelimizi veri setimizin bulunduğu klasörün içine kaydederek projemizi tamamladık.

```
Model name: Hist Gradient Boosting Regressor
Test accuracy score: 88.50%
Best parameters: {'learning_rate': 0.05, 'max_depth': 3, 'min_samples_leaf': 4}
```

REFERENCES

[1] https://www.kaggle.com/datasets/mirichoi0218/insurance

[2] https://www.kaggle.com