# Proje 1 : Medikal Maliyet Analizi

## Projenin Amacı

Verilen değişkenlere göre bir kişinin sağlık sigortasının yaklaşık ne kadar masraflı olacagını tahmin etmek.

## Model İçeriği

Her bir sağlık sigortası müşterisi ile ilişkilendirilen ve sigorta tarafından faturalanan tıbbi masraflarla ilişkili olduğu düşünülen değşikenlere sahip bir veri seti ile çalışılmıştır. Bahsedilen değişkenler aşağıdaki gibidir:

* Age (yaş): Sigorta sahibinin yaşı
* Sex (cinsiyet): Sigorta sahibinin cinsiyeti
* BMI (vücut kütle indeksi): Vücut kitle indeksi, vücut yapısının anlaşılmasını sağlayan boy ve kilo oranını kullanılarak hesaplanan bir indekstir. İdeal index değeri 18,5 ila 24,9 arasındadır.
* Children (çocuk sayısı): Ailedeki sağlık sigortası kapsamına alınan çocuk sayısı
* Smoker (sigara içme durumu): Sigorta sahibinin sigara içme durumu
* Region (bölge): Birleşik devletler içerisinde sigorta sahibinin bulunduğu konut bölgesi: northeast, southeast, southwest, northwest.
* Charges (ücretlendirme): Sigorta sahibine sigorta tarafından faturalanan tıbbi masraflar

## Model Açıklaması

*Bağımlı Değişken:* Charges

*Bağımsız Değişkenler:* Age, Sex, BMI, Children, Smoker, Region

*Kullanılan Regresyon Yöntemi:* CatBoostRegressor()

Veri seti biri hedef değişken olmak üzere toplam 7 kolondan oluşmaktadır ve 1338 satır içermektedir. Veri setinde null değere sahip hiçbir kolon bulunmamaktadır. Bağımsız değişkenler train setine göre normalize edilmiştir. Birçok regresyon yöntemi içerisinden CatBoostRegressor, regresyon metriklerine bakılarak ve 10-fold-cross-validation kullanılarak seçilmiştir. Model performansını arttırmak için Grid Search yöntemi ile Hiper-parametre Optimizasyonu yapılmıştır.

### Label Encoding ve One-Hot Encoding

Sex, Smoker ve Region değişkenleri numerik olmadıklarından Label Encoding ve One-Hot Encoding tekniklerine başvurulmuştur. Region değişkeninin değerleri sıralı değildir, örneğin southeast > southwest gibi bir ifade mantıken doğru olmayacaktır çünkü ikisi arasında matematiksel olarak ayırt edici bir kavram yoktur. Eğer Label Encoding kullanarak bu bölgelere 1,2,3,4 sayılarını vermeye kalkarsak model bu bölgeler arasında sayısal bir ilişki aramaya kalkabileceğinden bu bölgelerin her birinin 1-0 ile ifade edilen eşit ağırlıkta değişken kolonları olmaları modelin kolonları doğru anlaması açısından daha yararlıdır.Ama smoker ve sex kolonları için sadece iki tip değer barındırdığından One-Hot Encoding uygulamak ile Label Encoding uygulamak aynı şey olacaktır. Dolayısıyla Region değişkeni için One-Hot Encoding, Sex ve Smoker değişkenleri için Label Encoding kullanılmıştır.

### Regresyon Yöntemi

CatBoost, 2017 yılında Yandex adlı bir şirket tarafından geliştirilen nispeten yeni bir açık kaynaklı makine öğrenme algoritmasıdır. CatBoost, karar ağaçları ve gradient boosting teorisi üzerine kuruludur. Boosting ana fikri, birçok zayıf modeli (rastgele şanstan biraz daha iyi performans gösteren bir model) sırayla birleştirerek açgözlü arama yoluyla güçlü bir rekabetçi tahmin modeli oluşturmaktır. Gradient boosting algoritması sıre ile ilerlediğinden seçilen uygun ağaçlar önceki ağaçların hatalarından ders alacak ve dolayısıyla hataları azaltacaktır. Hata azaltma işlemi, hata artık küçültülemeyene kadar devam eder. Karar ağaçlarının büyüme prosedüründe, CatBoost genel gradient boosting modellerini izlemez. Bunun yerine, CatBoost kayıtsız ağaçlar yetiştirir, bu da ağaçların, tüm düğümlerin aynı seviyede olduğu, aynı tahmin ediciyi aynı koşulla test ettiği ormanlar yarattığı anlamına gelir. Kayıtsız ağaç prosedürü basitlik ve verimlilik sağlarken, ağaç yapısı optimal bir çözüm bulacak ve overfittingi önleyecek şekilde çalışır.

### Model Parametreleri

Grid Search metodu kullanılarak hiper-parametre optimizasyonu ile bulunan optimal parametreler aşağıda belirtilmiştir.

* Depth: 4
* l2\_leaf\_reg: 9
* Learning\_rate: 0.03

### Regresyon Metrikleri

* R2: 0.9018992656351004
* MAE: 2280.574321411652
* MSE: 15382987.874245934
* RMSE: 3922.1152295981738
* RMSE / MAE: 1.7197927700818911