**Kanada’daki Evlerin Regresyon Modelleri Kullanılarak Tahmin Edilmesi**

**Giriş:**

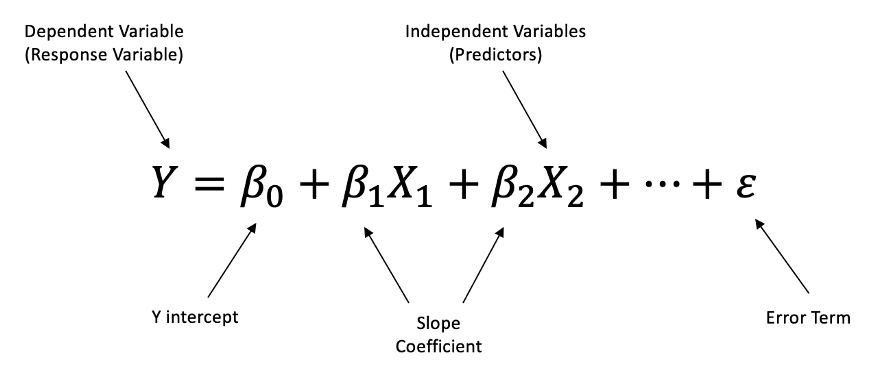
Regresyon nedir? Regresyon, istatistiksel bir analiz yöntemidir. Bir bağımlı değişken ile bir veya daha fazla bağımsız değişken arasındaki ilişkiyi modelleyerek bağımlı değişkeni tahmin etmek için kullanılır. Kesikli veri için uygun değildir, sürekli veri için tercih edilir.

**Modellerin Açıklanması:**

Bu ödevde 6 farklı regresyon modeli kullanılıyor. Sırasıyla bakalım:

**1- Linear Regression**

Ya da Türkçeye geçmiş haliyle Lineer Regresyon, adından da anlaşılabileceği üzere değişkenler arasındaki doğrusal bir ilişkiyi modellemeyi amaçlar. Ödevde kullandığımız veri için oldukça uygundur, zira bir evin fiyatı konumu, yatak odası sayısı gibi etkenlerle doğrusal ilişkilidir genelde. Formülü:



Dependent Variable: Tahmin ettiğimiz değişken.

Y intercept: Regresyon çizgisinin y eksenini kestiği değer (x = 0 değeri).

Independent Variables: Tahmin ederken kullandığımız değişken/ler.

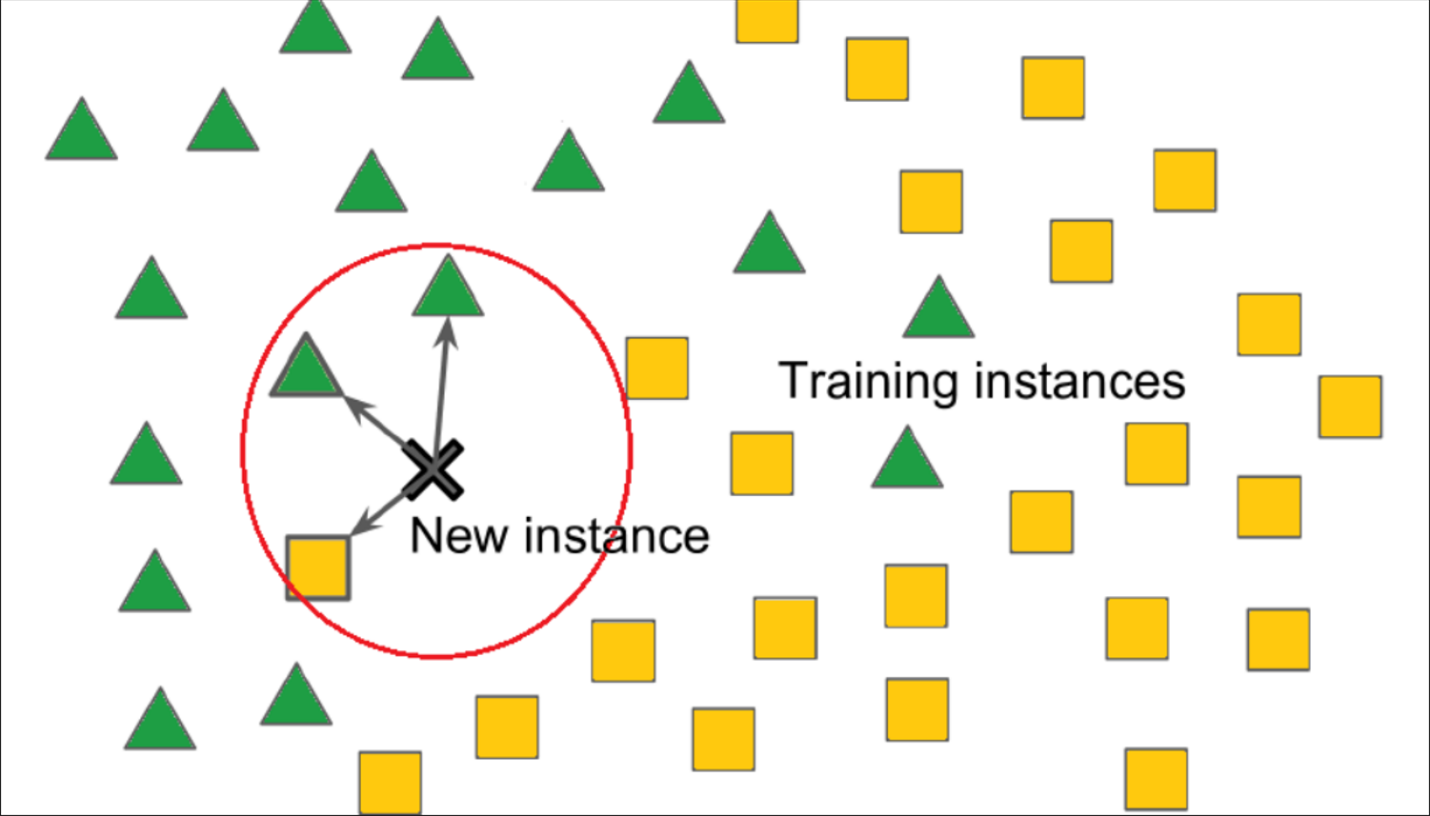
Slope Coefficient: Eğim katsayısı.

Error Term: Hata payı.

**2- kNN Regression**

kNN Regression, tıpkı kNN Classifier gibi k-Nearest Neighbors (k-En Yakın Komşu) mantığında çalışır. kNN Classifier’ın aksine kNN Regression k değerinin tek sayı olması gibi bir gereksinime sahip değildir. Veri setinin düzensiz veya gürültülü (noisy) olduğu durumlarda kullanışlıdır. Bu modelin dezavantajları olarak ise büyük ve çok boyutlu setlerde performans sorunu yaşatması ile uzaklık ölçüsü ve uygun k değeri gibi parametrelerin ayarlanması verilebilir.

kNN mantığını genel olarak tahmin edilecek veri noktasına belirli uzaklık fonksiyonları kullanarak en yakın k sayıdaki komşuyu bulup (regresyon için) bu komşuların bağımlı değişkenlerinin ortalamasını almak oluşturur. Bu sebeple yüksek boyutlu verilerde yavaş kalabilmektedir. kNN mantığı temel olarak şu şekilde gösterilebilir:



**3- Random Forest Regression**

Bu modelden başlayarak kullandığımız regresyon modelleri daha da karmaşık hal alıyor. Bu model, Random Forest algoritmasının regresyon versiyonudur. Bu model, birden çok karar ağacını bir araya getirip bu ağaçların tahminlerini birleştirerek bir regresyon modeli oluşturur.

Öncelikle veri seti rastgele alt kümelere bölünür. Her alt küme üzerinde ayrı bir karar ağacı oluşturulur. Karar ağaçları, veri setini özeliklerine göre böler ve her bölgeye bir regresyon tahmini yapar. Sonrasında bu tahminler birleştirilir, birleştirmek için genelde tahminlerin ortalaması alınır.

Doğası gereği karmaşık ve gürültülü setler için elverişli olabilir. Yüksek boyutlu veri setleri için de iyi bir seçenektir.

**4- Support Vector Regression**

Support Vector Machines (destek vektör makineleri) algoritmalarının regresyon versiyonudur. Özellikle doğrusal olmayan veriler için kullanışlıdır.

Önce eğitim verileri yüksek boyutlu bir uzaya haritalanır. Sonra hedef değişkenin etrafında bir tolerans aralığı belirlenir. Bu aralık içindeki vektörler destek vektör olarak kabul edilir. Bu destek vektörler sonrasında bu uzayda bir “hiper düzlem” oluşturur. Bu düzlem regresyon tahminlerinde kullanılır.

Daha önce açıklananın dışında, küçük ve overfittinge eğimli veriler üzerinde de kullanışlıdır.