****

**Kahramanmaraş**

**Sütçü İmam Üniversitesi**

Bilgisayar Mühendisliği

Makine Öğrenmesi Dersi

Proje Raporu

**Ülke Nüfuslarının Yıllara Göre**

**Büyüme Oranının Tahmin Edilmesi**

**Hazırlayanlar**

17110131006-Mehmet OKYAY

18110131505-Abdourazak ALİ EGUEH

Aralık-2021

**Projemizin Amacı:**

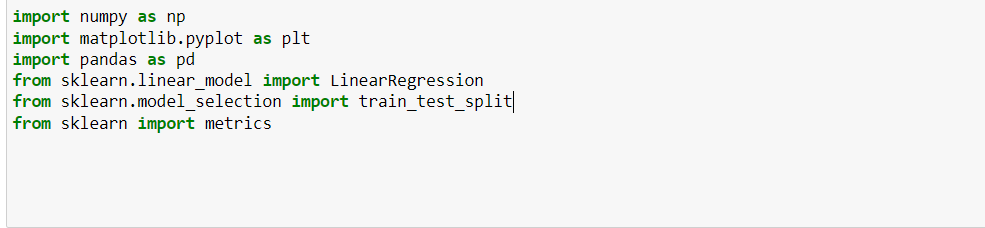
Projemizin amacı 1950-2011 yılları arasında ülkelerini nüfus bilgileri doğrultusunda belirlenmiş bir ülkenin, belirli bir yıla ait dair tahminlerde bulunmasını sağlamak.

**Veri Seti:**

Veri toplama adımı olarak Kaggle yer alan bir veri seti kullanılmıştır. Veri seti Birleşmiş Milletler(BM) tarafından 1950 yılında itibaren ülkelerin nüfus bilgilerinin yer aldığı bir veri setidir. Veri setinde 228 ülkenin 15109 adet kayıtlık nüfus verisi yer almaktadır.

**Verilerin Önişleme Adımları:**

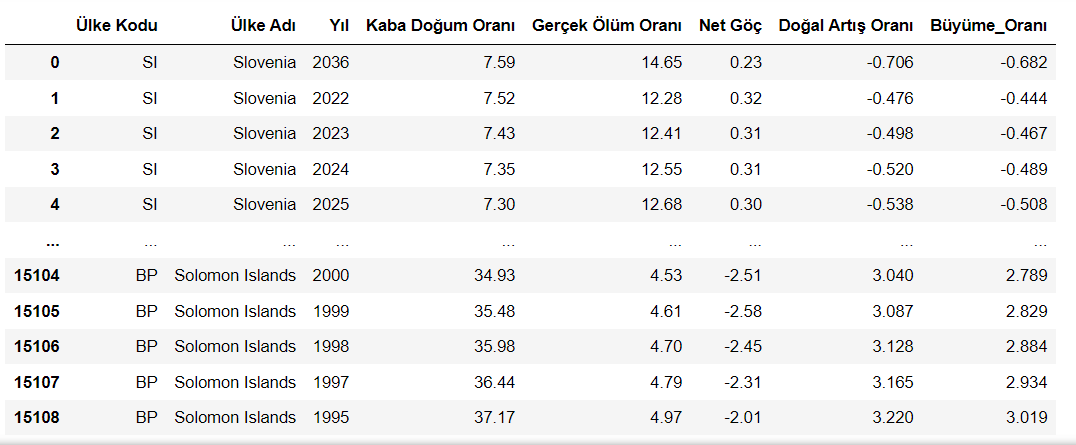
Verilerimizi elde edip bu verileri makine öğrenmesinde kullanılabilir hale getirmek için bir takım veri önişleme adımlarından geçirmek gerekmektedir. İlk olarak projemizde yer alan veri setini yüklemek için gerekli olan python kütüphanelerini projemize dahil ediyoruz.

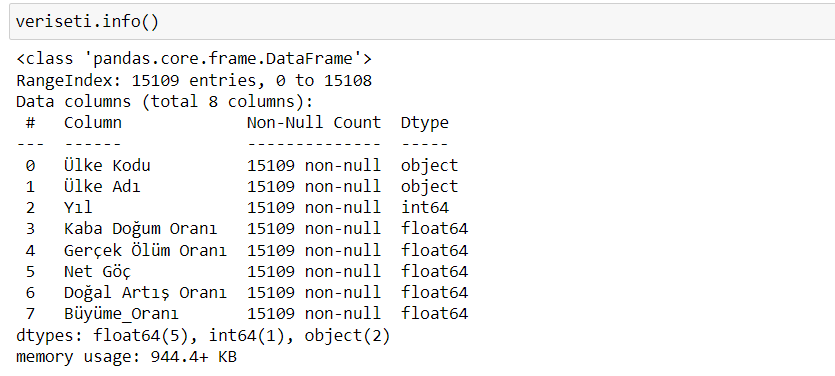


Veri setimizin sütun adlarının Türkçe karşıklarının yazılması sağlanmıştır.



Yüklediğimiz veri setine ait verilere örnek teşkil etmesi amacıyla ilk 5 ve son 5 değeri temsili olarak listelendi.

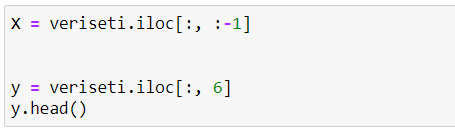


Veri setinde elde ettiğimiz verilere ait özellikler şöyledir. 8 adet nitelik olmaklar birlikte niteliklerden Ülke Kodu ve Ülke Adı object, yil integer ve diğerleri float türünde verilerdir. Verilerimizde herhangi bir boş değer bulunmamakla birlikte, veri setindeki toplam satır sayısı 15109 adettir.

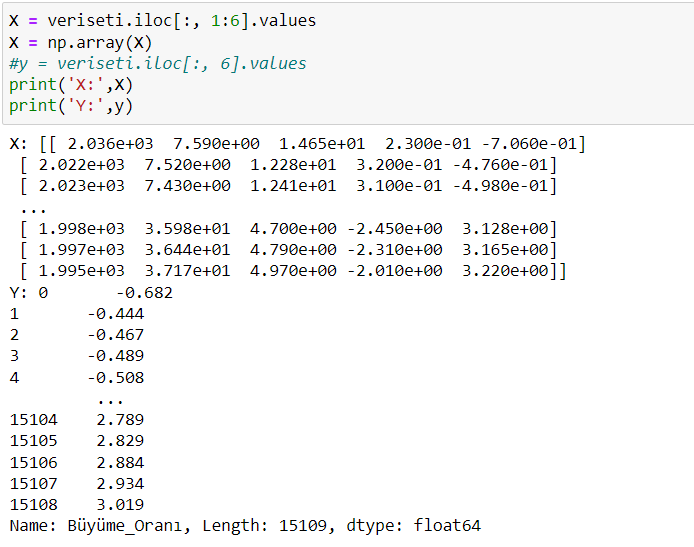
Veri setinde yer alan **ülke kodu** gereksiz olduğu düşünülerek Ülke Kodu sütunu silinmiştir.



Yüklediğimiz veri seti hakkında temel istatistiki verileri describe() fonksiyonu ile listelendi. 

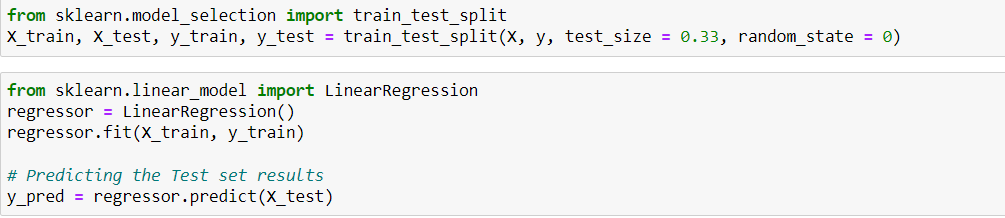
Verilerimizde kategorik veri olmaması boş değer olmaması verilerimizin eğitim modelinin oluşturulmasında yeterli olduğundan verilerimize ait bağımlı ve bağımsız değişkenler belirlenir. Veri setimizde X bağımlı değişkenimiz **büyüme oranı** iken geriye kalan **yıl, kaba doğum** **oranı**, **gerçek ölüm oranı**, **net göç** ve **doğal artış** **oranı** nitelikleri ise bağımsız değişkenler olarak belirlenmiştir.

Bağımlı ve bağımsız değişkenler ayrıldıktan sonra bir numpy verisi olarak ekrana yazırarak ayrışıp ayrışmadığı kontrol ediliyor.



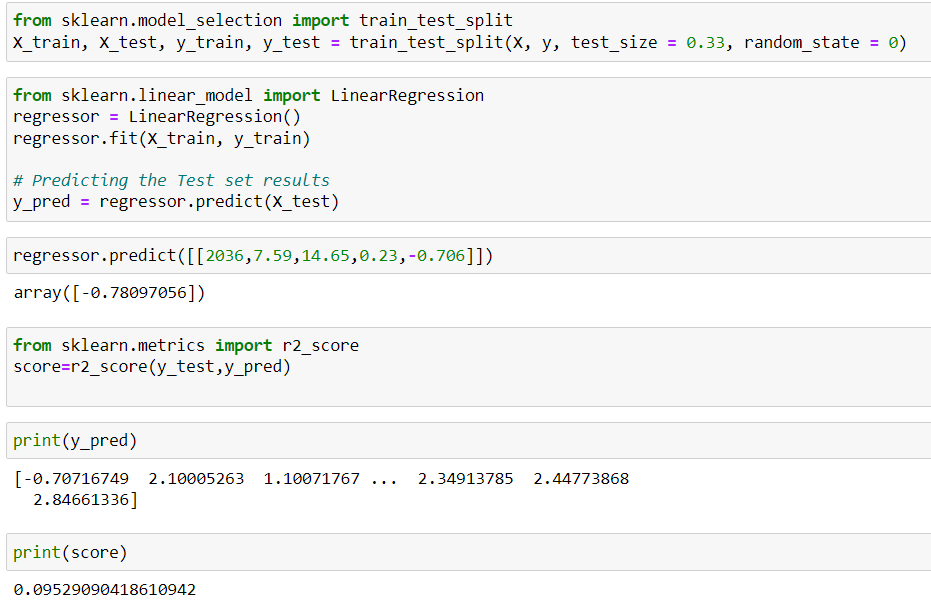
**Modelin Eğitilmesi ve Tahmin**:

Modelin eğitilmesi için gerekli olan python kütüphaneleri projeye eklenerek tran\_test\_split ile %33 lük bir oranla test ve eğitim eğitim verisi random olarak seçildi. R2 hata oranına bakılarak test\_size = 0,33 ve random\_state=0 olarak belirlendi. Bu değerler üzerinde değişikliğe gidildiğinde daha yüksek hata oranları görülmektedir.

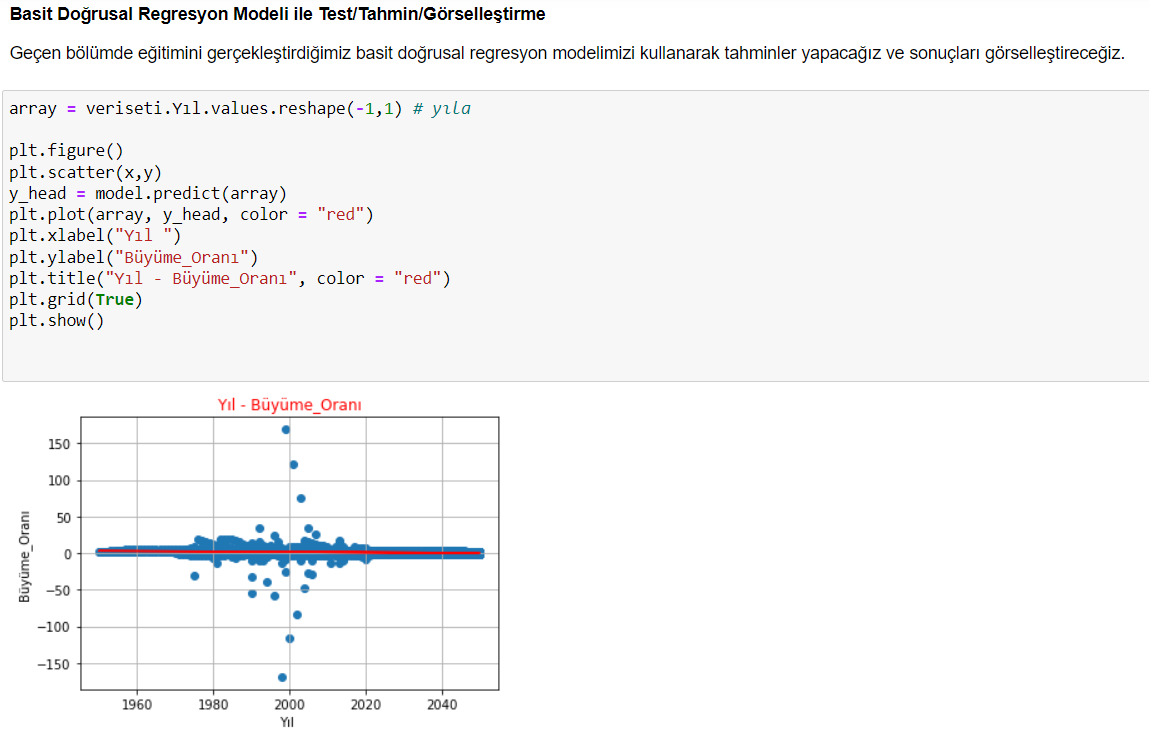


Tahmin işlemleri için lineer(doğrusal) regresyon modeli kullanıldı. Doğrusal regresyon, iki değişken arasındaki ilişkiye dayanarak birinin değerini diğerinden tahmin etmeyi sağlayan bir denklem (model) oluşturmayı içerir.

Doğrusal regresyon, bir dizi noktaya en uygun düz çizgiyi veya hiper düzlemi bulmak için kullanılmaktadır. Bir diğer ifadeyle doğrusal regresyon, en uygun düz çizgi (regresyon çizgisi olarak da bilinir ) kullanarak bağımlı değişken (Y) ile bir veya daha fazla bağımsız değişken (X) arasında bir ilişki kurar.



**Veriyi Görselleştirme**:



**Eğitim Modelinin Geliştirilmesi**:

R kare hata oranı değerlerine bakılarak eğitim modelinde bir test\_size, random\_state gibi parametrelerde değişikliklere gidilerek en az hata oranı tespit edilmeye çalışılmıştır.

