Veri Yoğun Uygulamalar (Spark) Ödevi

Konut Fiyatlarının Tahmini için Spark ile Regresyon Modeli Oluşturulması

Amaç ve Kapsam

Tanımlanan veri seti üzerinde konut fiyatlarının tahminlenmesi için Spark ML kütüphanesi kullanılarak PySpark ile bir regresyon modeli oluşturulacaktır.

Veri Kümesi

Veri kümesi: **California Housing Prices**

Veri hakkında detaylı bilgi için aşağıdaki sayfa incelenebilir. Veri, bu sayfadan indirilebilir.

https://www.kaggle.com/datasets/camnugent/california-housing-prices

Verinin bazı örnek alanları aşağıda verilmiştir:

- longitude
- latitude
- housing_median_age
- total_rooms
- ...

Makine öğrenmesi modeli ile tahminlenmesi hedeflenen veri alanı:

- median_house_value

Ayrıca, Spark regresyon dokümantasyonundan (Python için) yararlanılabilir:

https://spark.apache.org/docs/latest/ml-classification-regression.html

Ödev Kapsamında Hazırlanıp Teslim Edilecek Doküman

Ödev kapsamında geliştirilen PySpark kaynak kodlarını, her bir adımda yapılan işlemlerin açıklamalarını ve çıktılarını içeren rapor, tek bir PDF dokümanı olarak teslim edilecektir. Ödevde aşağıdaki adımlar gerçekleştirilecektir:

- Verilerin yüklenmesi
- Verilerin çeşitli Spark fonksiyonları kullanılarak incelenmesi
- Özniteliklerin seçimi ve verilerin makine öğrenmesi için hazırlanması,
- PySpark ile makine öğrenmesi modelinin oluşturulması
- Geliştirilen modelin performansının ölçümü