-README-

-Proje Açıklaması

Bu projede New York Airbnb Open Data ile çalışılmıştır. Bu data üzerinde çeşitli regresyon algoritmalarının performansını karşılaştırarak hangi algoritmanın bu veri seti için daha uygun olduğunu belirlemek amaçlanmaktadır. Kullanılan algoritmalar Lineer Regresyon, Ridge Regresyon, Random Forest ve Gradient Boosting'tir. Projede eksik verilerin nasıl işlendiği ve algoritmaların performansının nasıl değerlendirildiği hakkında detaylı bilgi sunuyoruz.

-Veri Seti

Veri Setinin Linki (https://www.kaggle.com/datasets/arianazmoudeh/airbnbopendata)

Veri seti, çeşitli özellikler içeren bir konaklama veritabanını temsil etmektedir. Eksik verilerin öncesi ve sonrası aşağıda gösterilmiştir:

Eksik Veriler Öncesi:
- `id` : 0
- `NAME` : 250
- `host id` : 0
- `host_identity_verified` : 289
- `host name` : 406
- `neighbourhood group` : 29
- `neighbourhood` : 16
- `lat` : 8
- `long` : 8
- `country` : 532
- `country code` : 131
- `instant_bookable` : 105
- `cancellation_policy` : 76
- `room type` : 0
- `Construction year` : 214
- `price` : 247
- `service fee` : 273
- `minimum nights` : 409
- `number of reviews` : 183
- `last review` : 15893
- `reviews per month` : 15879
- `review rate number` : 326
- `calculated host listings count` : 319
- `availability 365` : 448
- `house_rules` : 52131
- `license` : 102597

Eksik Veriler Sonrası:
- Tüm özellikler eksik verilerden arındırılmıştır.

-Verinin Eğitim ve Test Setlerine Bölünmesi

- Eğitim Seti: 82,079 örnek
- Test Seti: 20,520 örnek

-Algoritmalar ve Performans

Lineer Regresyon
- Çapraz Doğrulama Ortalama Skoru: -0.00013730127032702822
- Eğitim MSE: 109,872.42
- Test MSE: 109,178.29
- Eğitim R^2: 8.80e-05
- Test R^2: -0.00011940967061607743

Ridge Regresyon
- En İyi Parametreler: `alpha` = 10
- Çapraz Doğrulama Ortalama Skoru: -0.0001372680677439231
- Eğitim MSE: 109,872.42
- Test MSE: 109,178.29
- Eğitim R^2: 8.80e-05
- Test R^2: -0.00011938987466431072

Random Forest
- En İyi Parametreler: `max_depth` = 20, `min_samples_split` = 2, `n_estimators` = 200
- Çapraz Doğrulama Ortalama Skoru: 0.06684366084050475
- Eğitim MSE: 86,687.83
- Test MSE: 100,879.97
- Eğitim R^2: 0.21108309270141612
- Test R^2: 0.07589671982521573

Gradient Boosting
- En İyi Parametreler: `learning_rate` = 0.1, `max_depth` = 5, `n_estimators` = 200
- Çapraz Doğrulama Ortalama Skoru: 0.014429879611101116
- Eğitim MSE: 104,478.38
- Test MSE: 107,419.12
- Eğitim R^: 0.049177284008079525
- Test R^2: 0.01599538525221278

-Kümelenme Sonuçları ve Korelasyonlar

K-Means Kümelenme Sonuçları

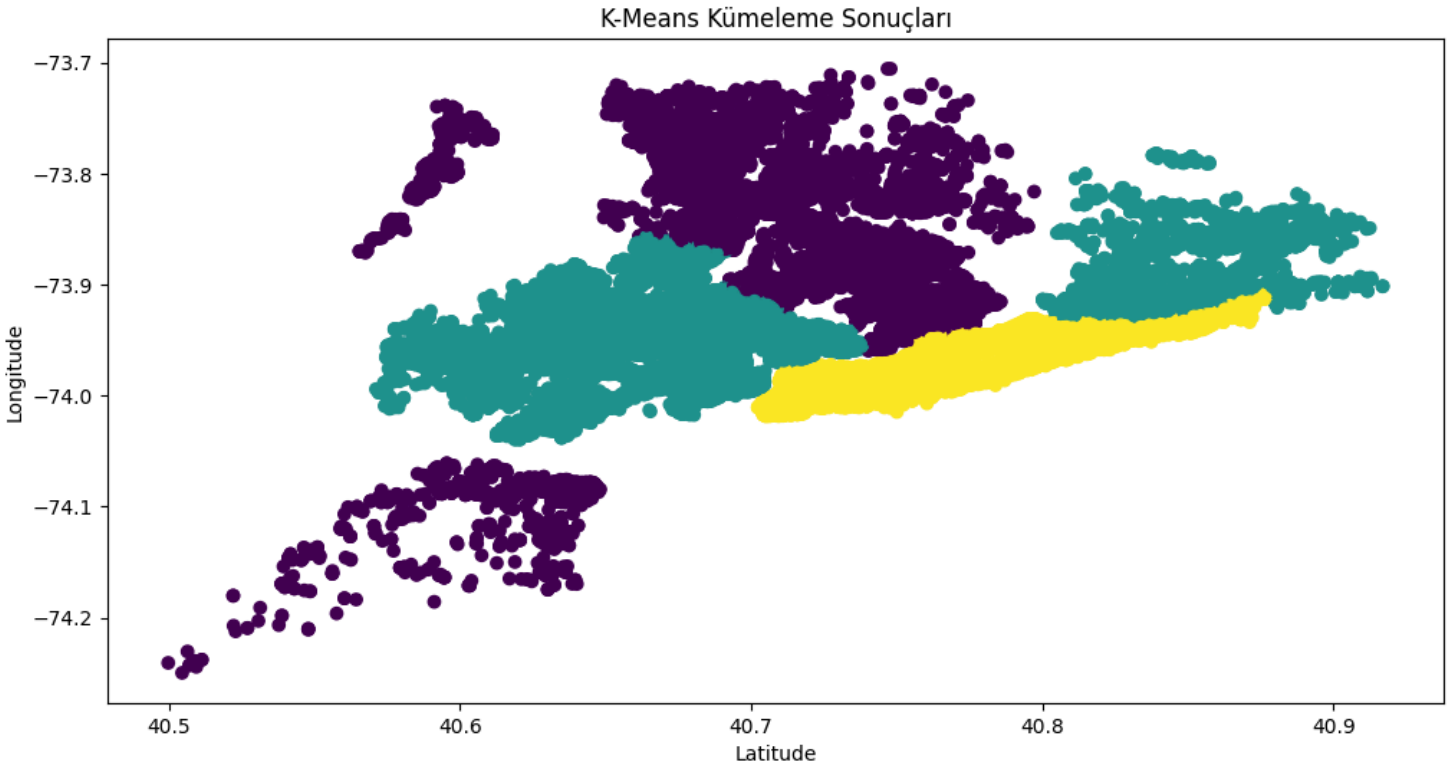
Bu görsel, K-Means algoritmasının çıktısını göstermektedir. K-Means, gözlemleri belirlenen k sayıda kümeye ayıran bir kümeleme algoritmasıdır. Görselde her bir gözlem, bulunduğu kümeye göre renklendirilmiştir ve kümelerin merkez noktaları belirtilmiştir. Aşağıdaki adımlar K-Means sürecinde uygulanmıştır:

Veriler öncelikle normalize edilmiştir.

Ardından, k değeri belirlenmiş ve K-Means algoritması uygulanmıştır.

Kümeler oluşturulurken, kümeler arasındaki mesafe minimize edilmiştir.

Bu görseldeki sonuçlar, verilerin hangi kümelere ayrıldığını ve bu kümelerin merkezlerinin nerede bulunduğunu göstermektedir.



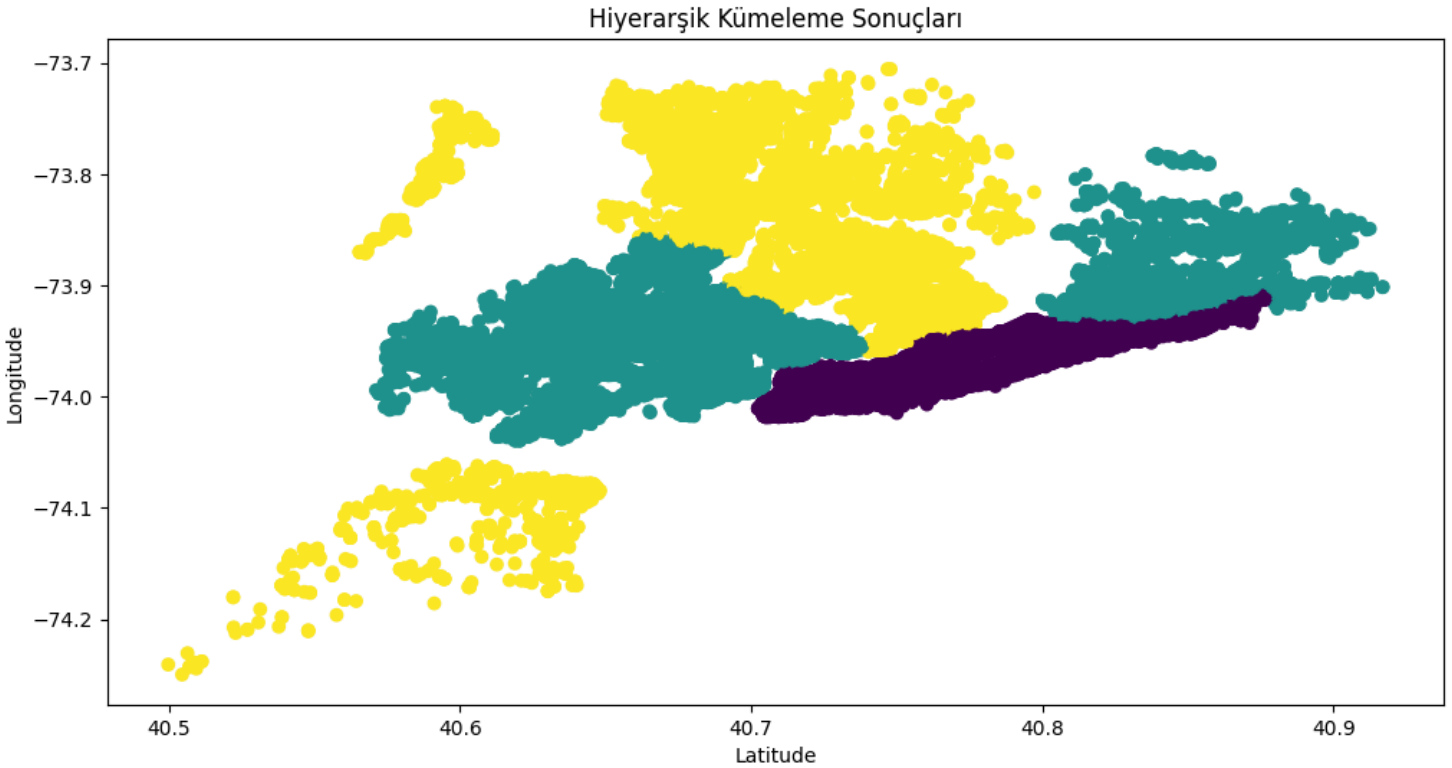
Hiyerarşik Kümelenme Sonuçları

Hiyerarşik kümeleme, kümelerin hiyerarşik bir yapıda oluşturulduğu bir algoritmadır. Bu görsel, verilerin birleşik veya ayrıştırılmış hiyerarşik kümeleme yöntemine göre gruplandığını göstermektedir. Hiyerarşik kümelemede:

Öncelikle her bir gözlem kendi kümesi olarak başlar.

Adım adım benzer kümeler birleştirilir (birleşimsel yöntem) ya da büyük kümeler daha küçük kümelere ayrılır (bölümleme yöntem).

Görselde, kümeler arasındaki bağlantı ve ilişkiler, dendrogram şeklinde gösterilmiştir. Bu dendrogram, kümelerin birleşme sıralamasını ve aralarındaki mesafeyi temsil eder.



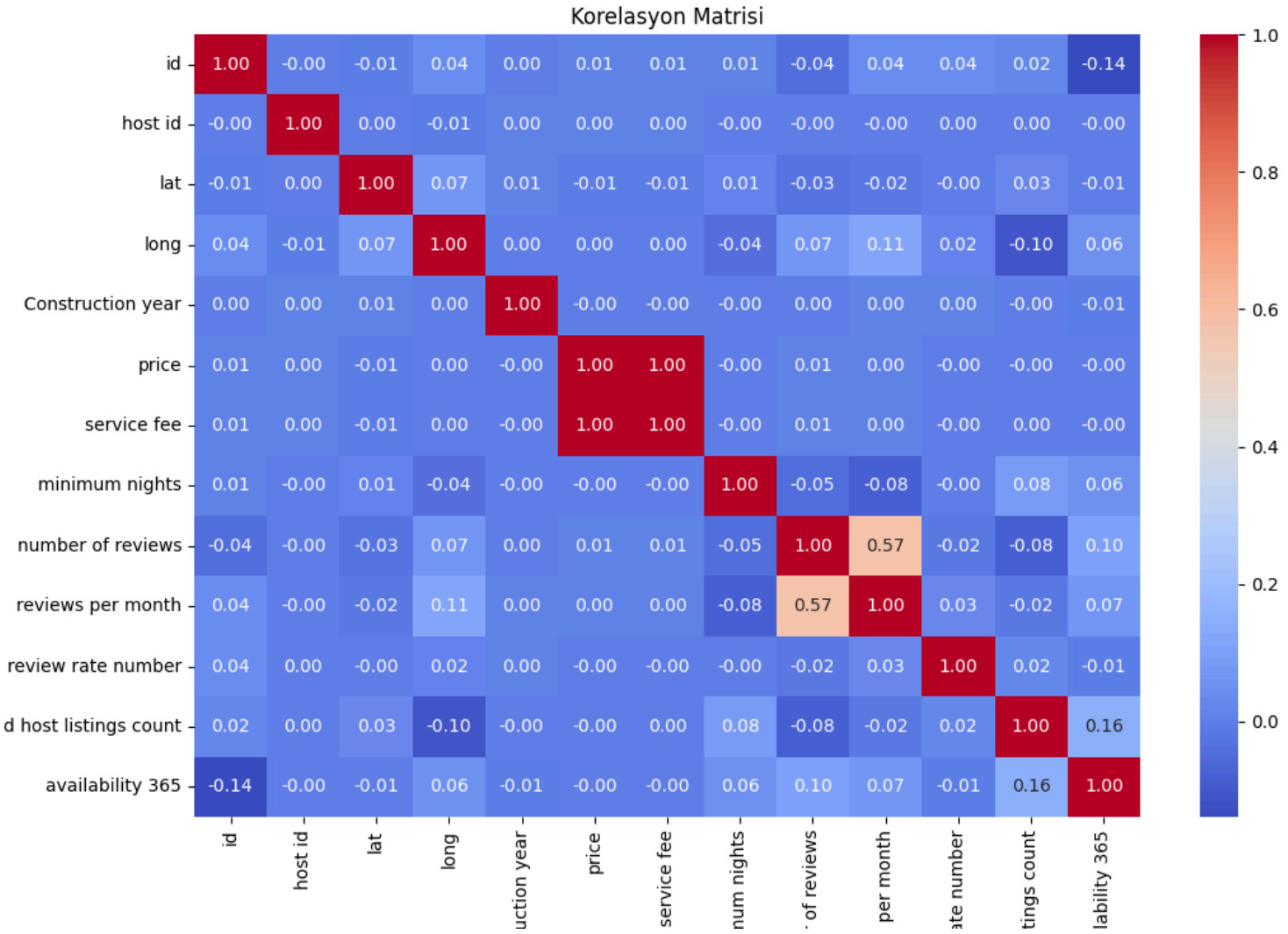
Korelasyon Matrisi

Bu görselde, veri setindeki değişkenler arasındaki korelasyonlar gösterilmiştir. Korelasyon, iki değişken arasındaki doğrusal ilişkiyi temsil eder ve genellikle Pearson korelasyon katsayısı ile ölçülür. Bu matris:

1 ile -1 arasında değerler alır, burada 1 güçlü pozitif bir ilişkiyi, -1 güçlü negatif bir ilişkiyi gösterir.

0'a yakın değerler, iki değişken arasında anlamlı bir ilişkinin olmadığını gösterir.

Görselde, her hücre bir değişken çiftinin korelasyonunu temsil eder ve renk kodlaması, korelasyon katsayılarının görsel olarak kolayca ayırt edilmesine yardımcı olur.



-Sonuç ve Karşılaştırma

Performans Karşılaştırması:

- Lineer Regresyon ve Ridge Regresyon:
Her iki model de benzer sonuçlar elde etmiş ve test MSE ve R^2 değerleri oldukça düşük olmuştur. Bu modeller, veri setinin doğrusal ilişkileri modelleme açısından uygun olmadığını göstermektedir.

- Random Forest:
Daha iyi performans göstermiştir; eğitim ve test MSE değerleri diğer modellerden daha düşük olup, R^2 değerleri daha yüksektir. Bu model, daha karmaşık ve non-lineer ilişkileri öğrenme kapasitesine sahiptir.

- Gradient Boosting:
Orta düzeyde bir performans göstermiştir; test MSE ve R^2 değerleri Random Forest'a göre daha düşük, ancak yine de diğer regresyon modellerinden daha iyi sonuçlar elde edilmiştir.

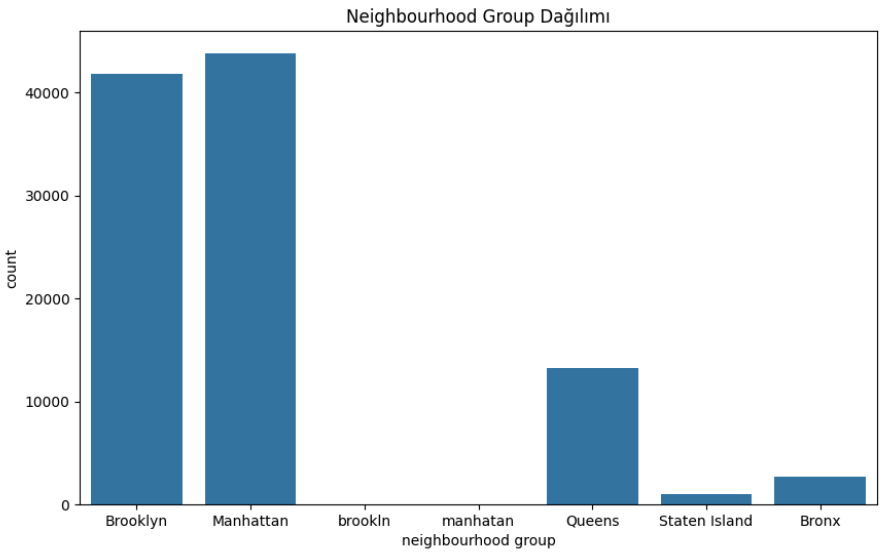
Sonuç:
- Random Forest, veri setinin karmaşıklığını ve non-lineer ilişkileri en iyi şekilde yakalayan algoritma olarak öne çıkmaktadır. Bu, algoritmanın güçlü ağaç tabanlı öğrenme yöntemleri ve en iyi parametre ayarları sayesinde mümkün olmuştur.

- Lineer Regresyon ve Ridge Regresyon, doğrusal modelleme için uygun olmakla birlikte, bu veri setinin özellikleri ve karmaşıklığı göz önüne alındığında yeterince iyi performans gösterememiştir.

Veri Analizleri:
Bu veri setinden örnek olarak çıkardığım iki analiz şunlardır:

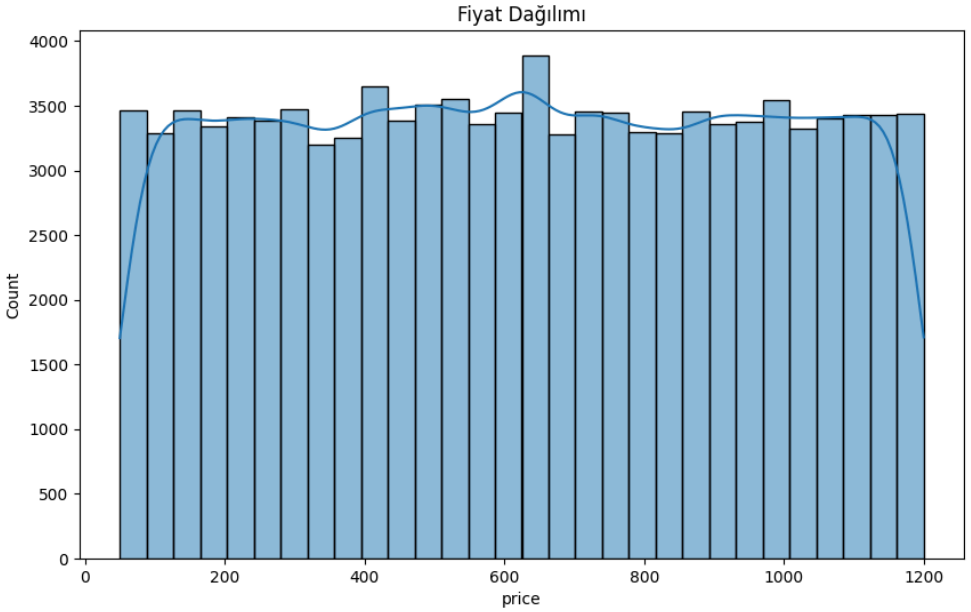
1- Eyaletlere Göre Airbnb Sayısı:

İlk analiz, eyaletler arasındaki Airbnb sayısını karşılaştırmaktadır. Bu analiz, farklı eyaletlerde bulunan Airbnb'lerin toplam sayısını incelemektedir.



2- New York Bölgesindeki Airbnb'lerin Fiyatlarına Göre Dağılımı:

İkinci analiz, New York bölgesindeki Airbnb'lerin fiyatlarına göre sayılarını karşılaştırmaktadır. Bu analiz, New York bölgesinde bulunan Airbnb'lerin fiyat aralıklarına göre dağılımını göstermektedir.



-Öğrenilenler

1. Eksik Verilerin İşlenmesi:
Eksik verilerin nasıl temizleneceği ve veri setinin eksik verilerden nasıl arındırılacağı konusunda önemli bilgiler edindik. Eksik verilerin işlenmesi, model performansını doğrudan etkileyen bir faktördür.

2. Algoritma Performansı:
Farklı regresyon algoritmalarının performansını değerlendirme konusunda tecrübe kazandık. Özellikle Random Forest ve Gradient Boosting gibi karmaşık modellerin, doğrusal regresyon modellerine göre daha iyi performans gösterdiğini gözlemledik.

3. Model Seçimi:
Doğru model seçiminin, veri setinin özelliklerine bağlı olarak değiştiğini anladık. Random Forest, veri setindeki non-lineer ilişkileri en iyi şekilde modelleyerek en yüksek performansı sağladı.

4. Model Değerlendirme:
Model performansını değerlendirmenin ve sonuçları yorumlamanın önemini kavradık. MSE ve R^2 gibi metriklerin model değerlendirme sürecindeki rolünü vurguladık.

5. Parametre Ayarları:
Algoritmaların performansını optimize etmek için doğru parametre ayarlarının nasıl yapılacağı hakkında bilgi edindik. Parametre ayarlarının modelin başarısını arttırmakta kritik bir nokta olduğunu anladık.

6. Denetimli Öğrenme:
Denetimli öğrenme algoritmalarının çalışma prensipleri ve farklı türde veri setleri için uygunlukları hakkında daha derin bir anlayış kazandık.

Proje Sahipleri:

Eylül KILIÇ mail:eylulkilic977@gmail.com
Efe TUNÇ mail:efeahmet259@gmail.com