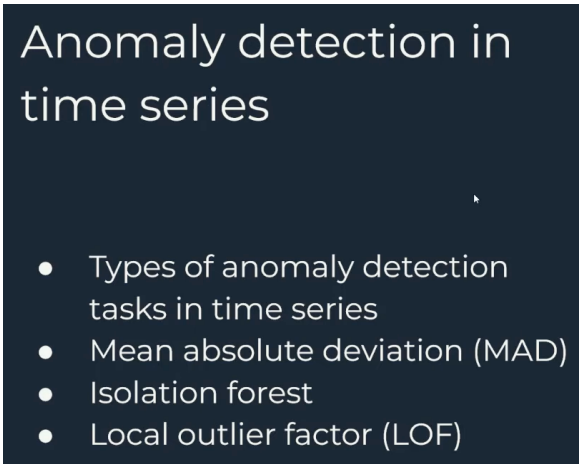
✅

Anomaly detection in time series with Python (Video)

Issues	Anomaly Detection, Isolation Forest, Local Outlier Factor, Median Absolute Deviation
Link	https://www.youtube.com/watch?v=qy41dXGbAxY
Status	Finished
Reason	
Starting Date	@December 23, 2023

▼ Agenda



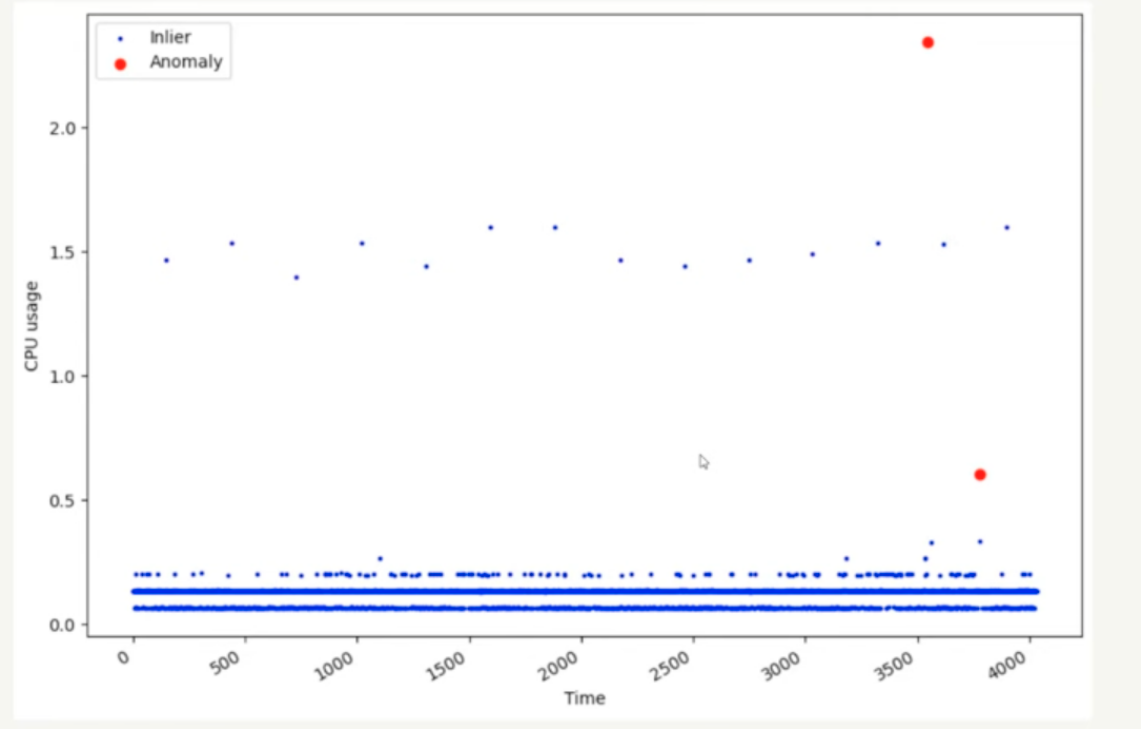
▼ Neden Anomaly Detection Yapıyoruz?

- Diyelim ki Amazon ziyaretçi sayıları bir anda 0’a indiyse orada anormal bir durum var demektir. Belki web site çöktü, hemen müdahale edilmesi gerekiyor.
- İkinci olarak da normal tahmin modellerinde outlier değerleri history datanda istemiyorsan onları tespit etmek için de anomaly detection yapıyoruz. Onları tutmalı mıyız, belki tamamen atacağız, belki normalize edeceğiz.

▼ 2 tür anomaly detection taskı var.

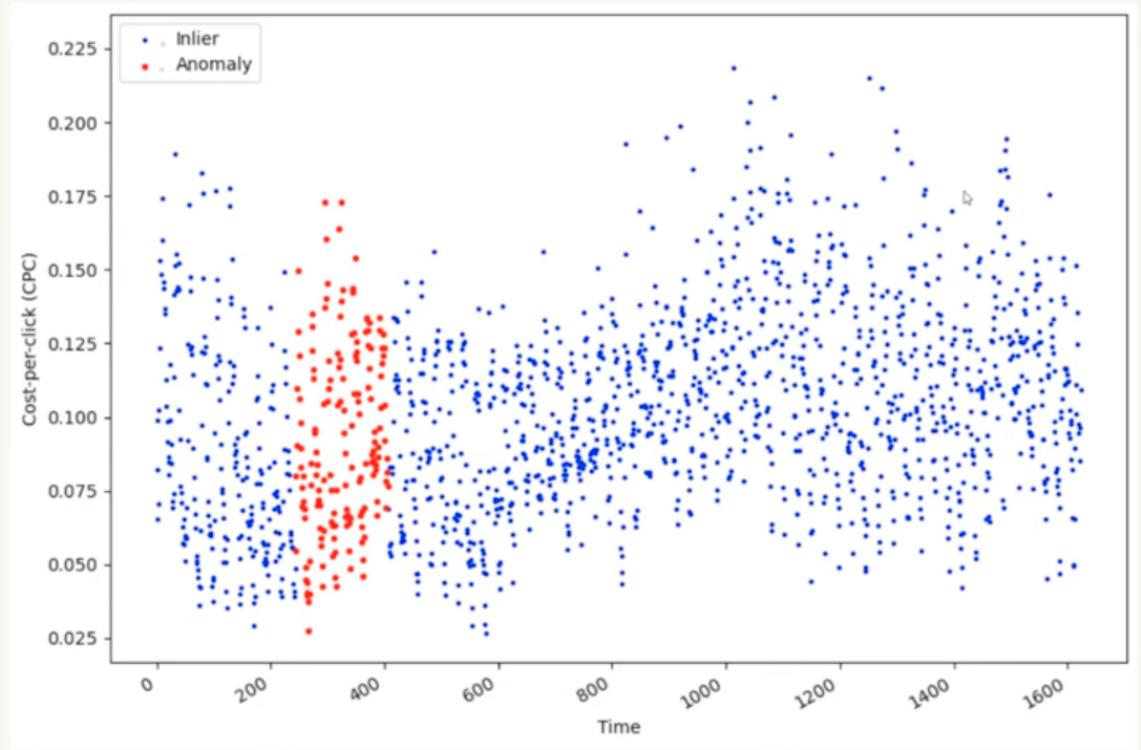
▼ 1. Point-wise anomaly detection

Bu derste de üzerinde çalışacağımız senaryo: bir AWS EC2’nin CPU kullanımıdır. Buradaki 2 kırmızı nokta isolated anomaliler.



▼ 2. Pattern-wise anomaly detection

Anormal bir model oluşturan bir dizi noktayı tanımlamaya çalışıyoruz. Örnek olarak bir reklamın tıklama başına maliyetini izliyoruz. Buradaki kırmızı nokta serisi anormal olarak değerlendiriliyor.

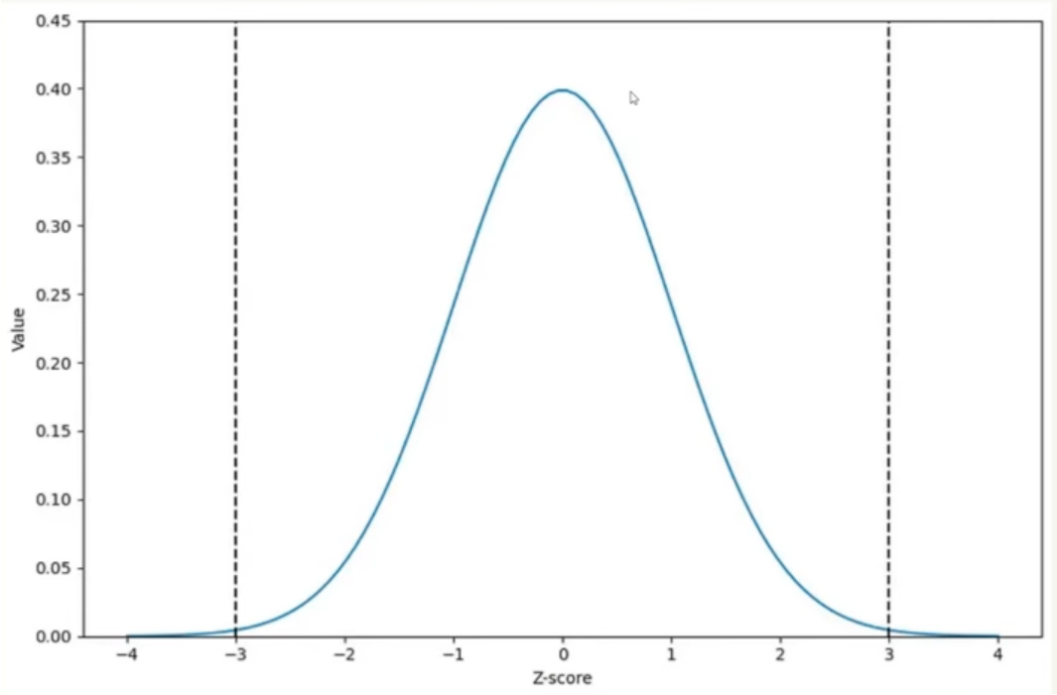


▼ Mean absolute deviation (MAD) (Baseline method)

Veriler normal şekilde dağıldığında, her taildeki (kuyruktaki) noktaların aykırı değerler olduğu sonucuna varabiliriz.

Z score 3 veya 3.5'un üzerindeyse bunu outlier olarak kabul ediyoruz. Aynısı negatif değerler için de geçerli.

$$Z = \frac{x - \mu}{\sigma}$$

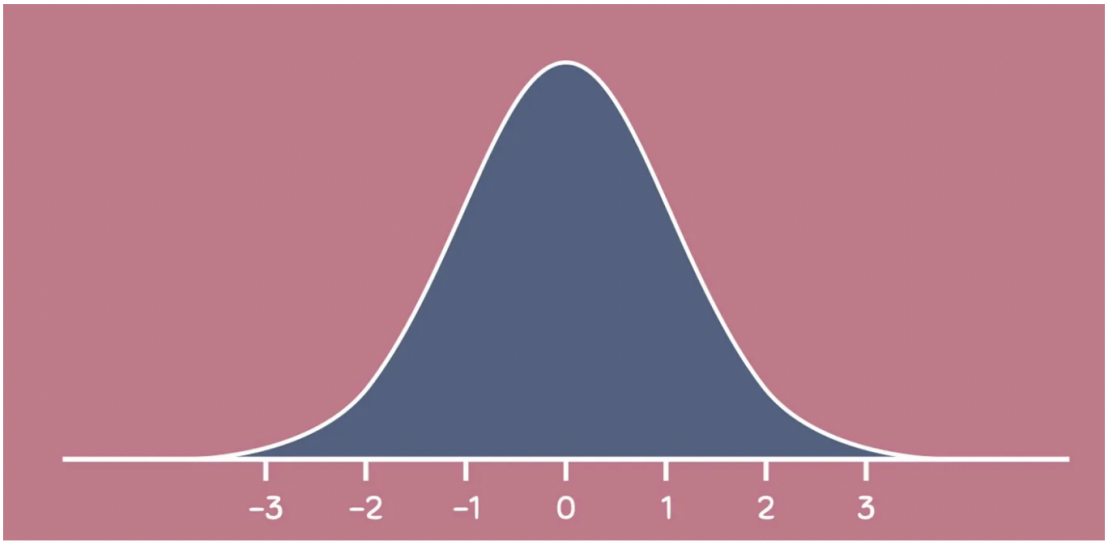


▼ Z-Skor Neydi

Normal Dağılım , Z-Score Standardizasyon ve Normalizasyon

Standart Normal Dağılım'ın her zaman orta noktası (μ) 0'dır ve aralıklar birer birer artar.

Yatay eksendeki her sayı bir z-puanına karşılık gelir. z -puanı bize bir gözlemin ortalamadan(μ) kaç adet standart sapma uzak olduğunu göstermektedir.



Standart Normal Dağılım

Fakat burada şöyle bir durum var. Outlier varken senin meanin (u) de etkileniyor. Dolayısıyla da z score değerin etkileniyor.

Buradan da örneğe bakılabilir.

Mean absolute deviation (MAD) review

▼ Median absolute deviation (MAD)

Outlier'dan etkilenmeyecek bir hesaplama şekli bulmamız gerekiyor. Bu yüzden mean yerine **median** kullanıyoruz. *(Mean'den ne kadar uzaktaydı noktalar yerine medianından ne kadar uzakta diye bakıyoruz ve bunun da medianını alıyoruz.)*

For a univariate data set X_1, X_2, ..., X_n, the MAD is defined as the median of the absolute deviations from the data's median $\tilde{X} = \text{median}(X)$:

$$\text{MAD} = \text{median}(|X_i - \tilde{X}|)$$

Z scoru artık bu şekilde hesaplıyoruz. *(Artık noktamın ortalamadan kaç standart sapma uzaklığında olduğuna değil medyandan kaç MAD uzaklığında olduğuna bakıyorum.)*

$$M_i = \frac{0.6745(x_i - \tilde{x})}{MAD}$$

Neden 0.6745 ile çarpıyoruz? Çünkü z-score median absolute value'yu (MAD) kullanıyor. Bu her zaman standart sapmadan daha küçüktür. (Asıl formülde standart sapma vardı) Z-score'una yaklaşabilmek için payı 1'den küçük bir sayıyla çarpıyoruz.

- MAD'ı hesaplarken tüm data için ortak olarak hesaplamış oluyoruz. Ama Z skor veri noktasının skoru oluyor.

▼ Bu metodu kullanırken dikkat etmemiz gerekenler

Be careful!
- Robust Z-score works only if the data is close to a normal distribution
- The MAD is not equal to 0 (happens when more than 50% of the data has the same value)

▼ Isolation Forest

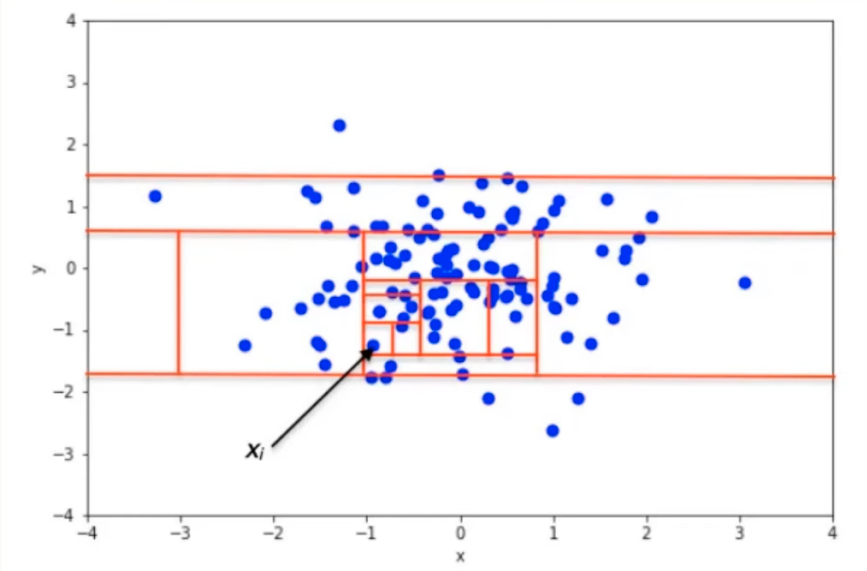
Isolation forest

Tree-based algorithm to detect outliers

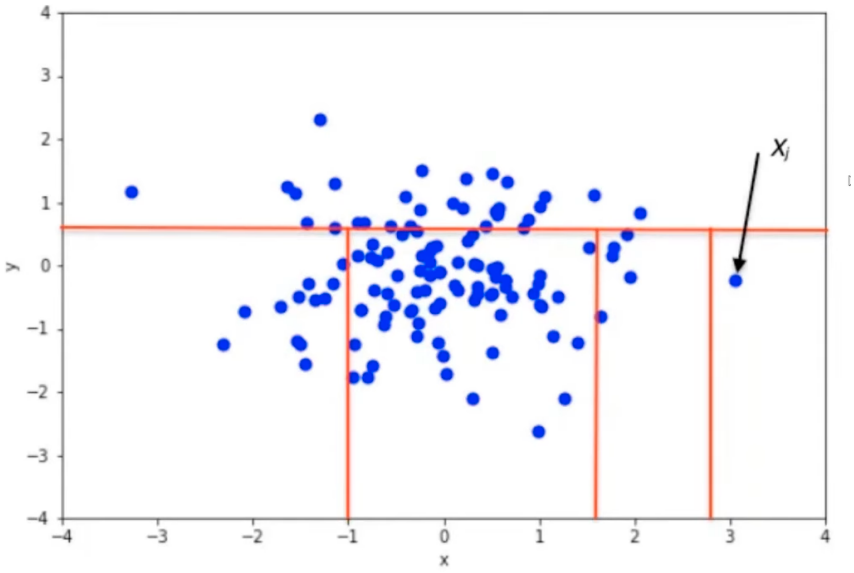
Partitions the data to isolate points
- Many partitions - means the point is an inlier
- Few partitions - means the point is an outlier

Algoritma rastgele bir attribute seçerek başlar ve ardından bu attribute için maksimum ve minimum değerler arasında rastgele bir bölünmüş değer seçersiniz ve böylece algoritma veri kümesindeki her noktayı izole edene kadar bölümleme birçok kez yapılır. Ve bunun arkasındaki genel fikir şu; eğer belirli bir noktayı izole etmek için çok sayıda bölüme ihtiyaç duyulursa bu, noktanın bir inlier olduğu anlamına gelir. Bununla birlikte, onu izole etmek için birkaç bölüme ihtiyacınız varsa, o zaman bu noktanın bir outlier olduğu anlamına gelir.

Burada xi'yi ayırt etmek için çok fazla bölümlemeye ihtiyaç duyuyoruz. O yüzden bu nokta inlier.



Burada ise xj'yi ayırt etmek için sadece 4 bölmeye ihtiyaç duyuyoruz. O yüzden bu nokta outlier.



▼ Local Outlier Factor (LOF)

Local outlier factor (LOF)

Unsupervised method for anomaly detection

Intuition: compare the local density of a point to that of its neighbors

If the density is smaller, then the point is isolated, so it's an outlier

Local outlier factor (LOF)

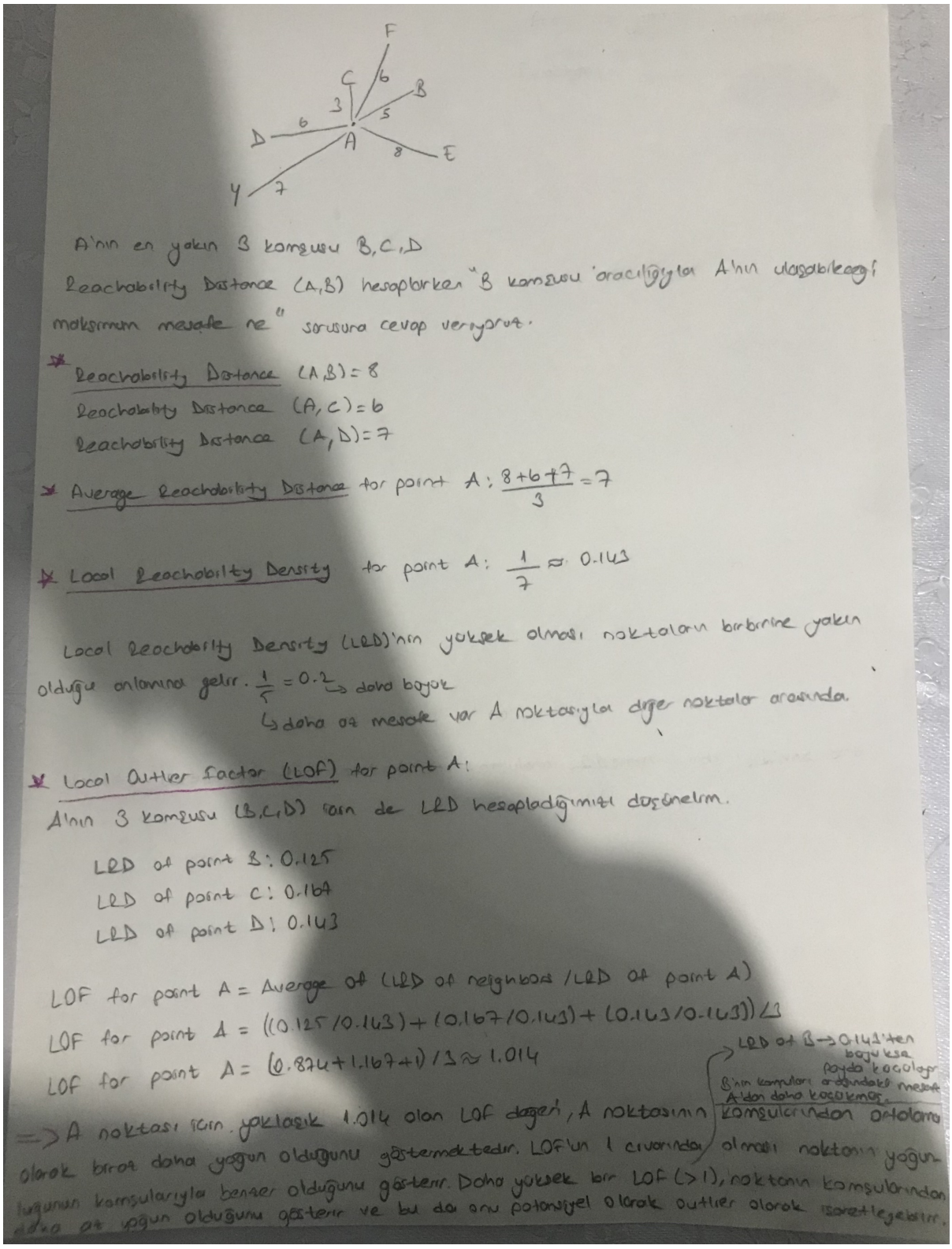
Unsupervised method for anomaly detection

Intuition: compare the local density of a point to that of its neighbors

If the density is smaller, then the point is isolated, so it's an outlier

✅ If LOF is close to 1 or smaller than 1: it's an inlier

❌ If LOF is larger than 1: it's an outlier



LOF'un 1 civarında olması noktanın yoğunluğunun komşularıyla benzer olduğunu gösterir. Daha yüksek bir LOF (>1), noktanın komşularından daha az yoğun

olduğunu gösterir ve bu da onu potansiyel olarak aykırı değer olarak işaretleyebilir.