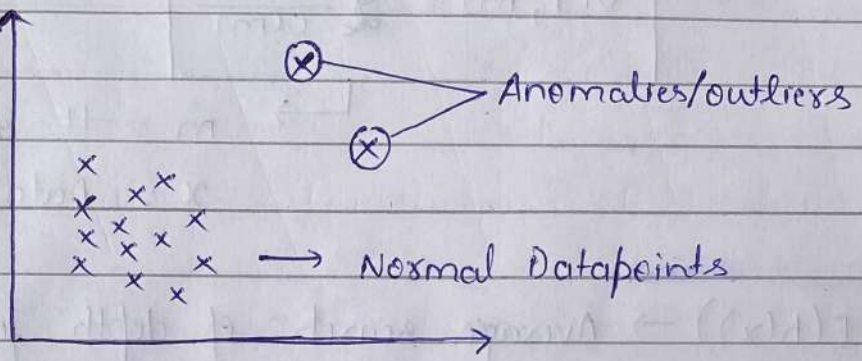
Anomaly Detection

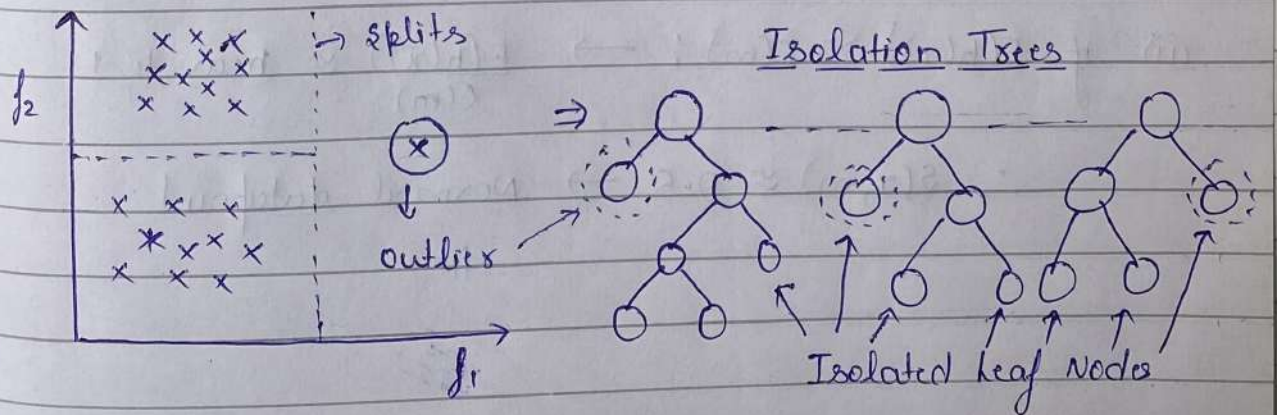
→ It is a technique in machine learning to identify the datapoints that are different from other datapoints.



Note: Anomalies in a dataset can be identify from the following:

(i) Isolation Forest
(ii) DBScan
(iii) Local Outlier Factor

* Isolation Forest: → Decision Trees



Note: For datapoints we are getting quickly isolated leaf nodes or when depth is small then the probability of those points to be anomalies are high.

→ Anomaly Score : → use to find whether an isolated point is an anomaly or not.

$$S(x, m) = 2^{-\frac{E(h(x))}{c(m)}}$$

↳ m = # of data points
x = Data point

E(h(x)) → Average search of depth for x from the isolated tree.

c(m) → Average depth of all the data points.

- Case :

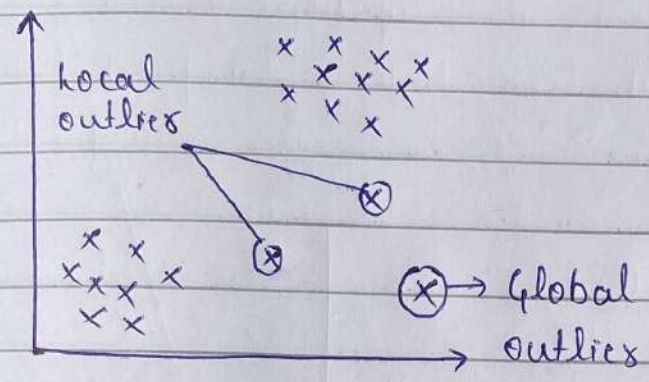
(i) if $E(h(x)) << c(m)$: → $\frac{E(h(x))}{c(m)} \approx 0$

∴ $S(x, m) \approx 1$ → Anomaly Score → outlier

(ii) if $E(h(x)) >> c(m)$: → $\frac{E(h(x))}{c(m)}$ ≈ near to 1

∴ $S(x, m) \approx 0.5$ → Normal datapoint

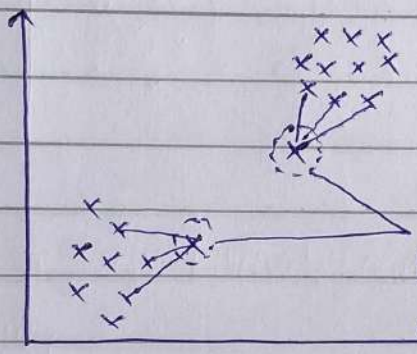
* Local Outlier Factor :



LOF

(i) local outlier

(ii) Global outlier



let K = 3

local density of these points is smaller than neighbouring 3 datapoints.

Note : It consider a datapoint as an outlier / anomaly by comparing the local density of that datapoint to the local densities of its neighbors.

- If sample's / datapoint's local density is smaller then it is an outlier.