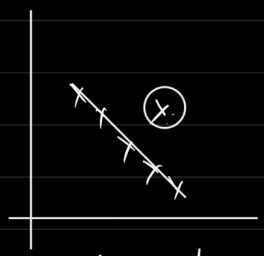
Anomaly detection (To detect outliers)

⇓

Abnormality

→ Anomaly detection algos are Unsupervised.



→ sometimes to detect outliers traditional methods like Box plot, violin plot fails.

Why?

→ In few case Anomaly plays an important role.

Ex1

transaction amount	Acc no
1000	—
2000	—
1500	—
3000	—
2 Lakhs	—

→ This customer being outlier, holds an important role. ⟹ Bank will target this customer for the loan.

∴ Outliers sometimes makes sense.

Ex2 → Fraud transaction.

tr_1 – Delhi
tr_2 – Delhi
tr_3 – Delhi
} Normal

$tran_4$ – USA ⟶ you receive a msg ⟹ anomaly.

Ex3 → When you login a gmail or any portal.

Ex4 VK. → 5, 10, 5, 4, 3, 5, 110 → This outlier but important.

Ex5 Cancer prediction ⇒ Normally people don't have Cancer.

Ex6 → Fraud IP.

* Detection of Outliers

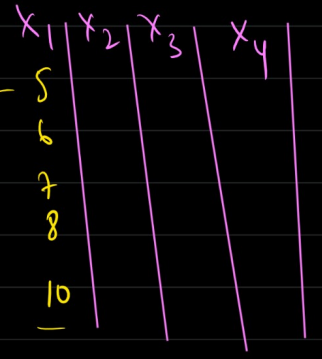
1. Isolation forest Anomaly detection
2. DBSCAN
3. Local Outlier Factor Anomaly detection.

1. Isolation forest (USL algorithms)

⇓

⇒ multiple Isolation trees

→ Isolation tree is like a decision tree

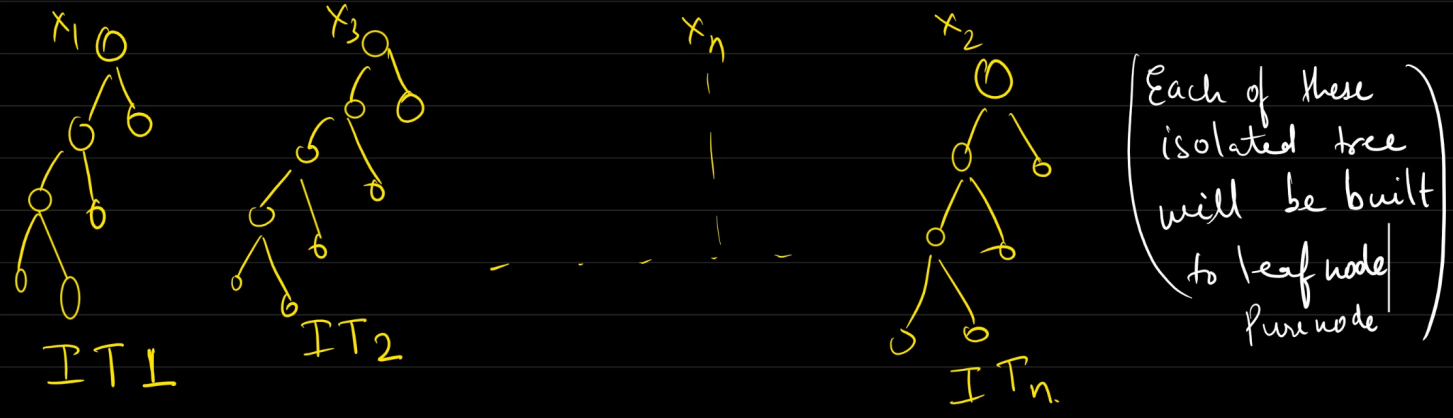


Construction of Isolation tree

→ Select a feature randomly X_1

→ Randomly choose a split value within the range of X_1

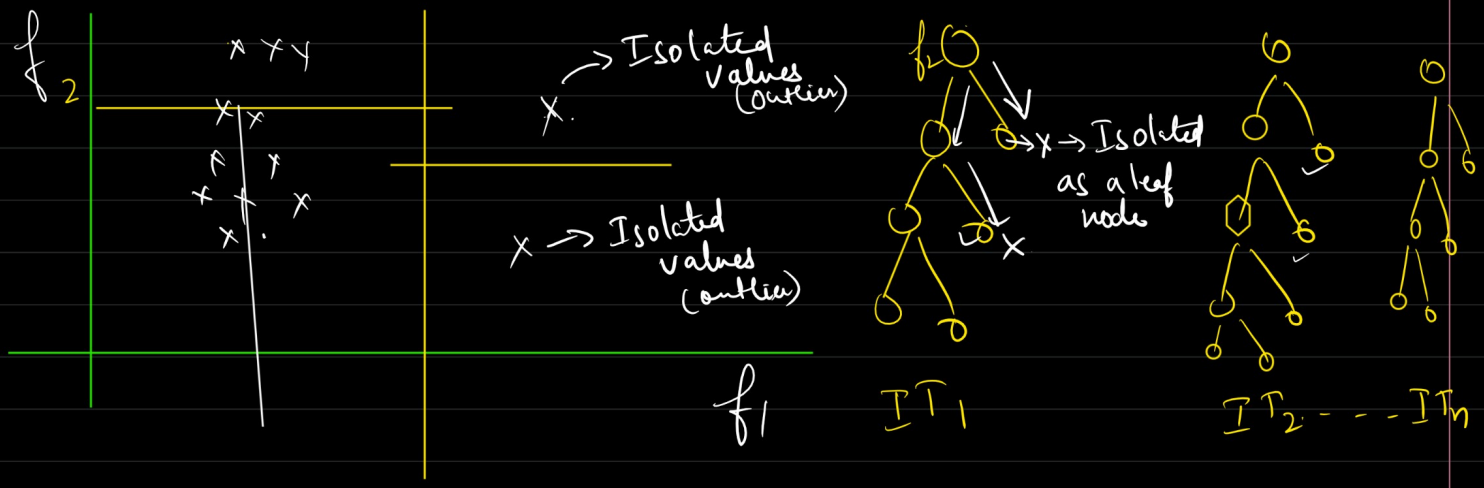
→ Repeat the process recursively to build the tree.



(Each of these isolated tree will be built to leaf node | Pure node)

* How you will get outlier?

→ Anomaly | outlier due to their distinctiveness, tend to end up in leaf nodes at shorte path.



* Since outliers are different from normal dp, during construction of isolated trees it will be isolated as a seperate leaf node.

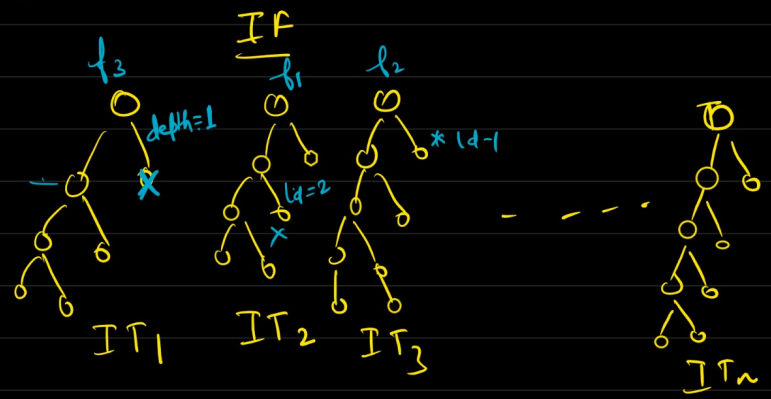
* **Anomaly score**

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

where m is total no of datapoints
x → datapoint for which you want to check anomaly score.

$E(h(x))$ = Average search depth of x in all of the isolated tree.
↓ for one dp

$C(m)$ = Average depth of all the dp's in all Isolation tree



$E(h(x)) \ll C(m)$ Since C(m) is avg depth for all the dp's so C(m) will be far greater than h(x)

$S(x,m) := 2^{\frac{-E(h(x))}{C(m)}}$ → will be very small value

$S(x,m) \approx 1 \Rightarrow$ Anomaly score $\Rightarrow$ outlier

Generally If $S(x,m) \geq 0.5 \Rightarrow$ outlier.

↑ threshold

$E(h(x) >> C(m) \Rightarrow S(x,m) \approx 0.5 \Rightarrow$ Normal dp's

< 0.5