* Iterative Imputer | MICE

Used when data is missing at Random (MAR)

Advantage
↳ quite accurate.

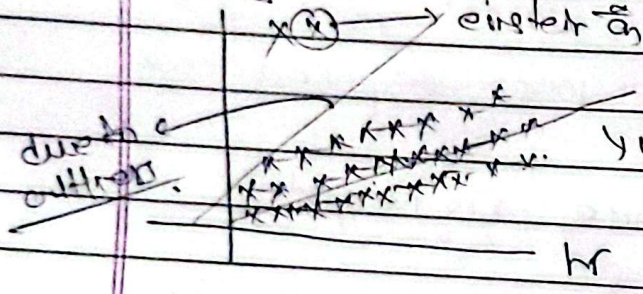
Disadvantage
↳ slow
↳ memory problem

Main Motive
↳ predicting missing value using MICE

* Outliers

What are outliers?

↳ शार्क जिका होता → 99% exam में but still crying. marks



Why outlier dangerous?

eg:
Age → 300

But:

outliers are helpful in Anomaly detection.
like credit card anomaly detection.
↳ should be used in anomaly detection

outlier detection → simple
But
what to do with outliers is very tedious job

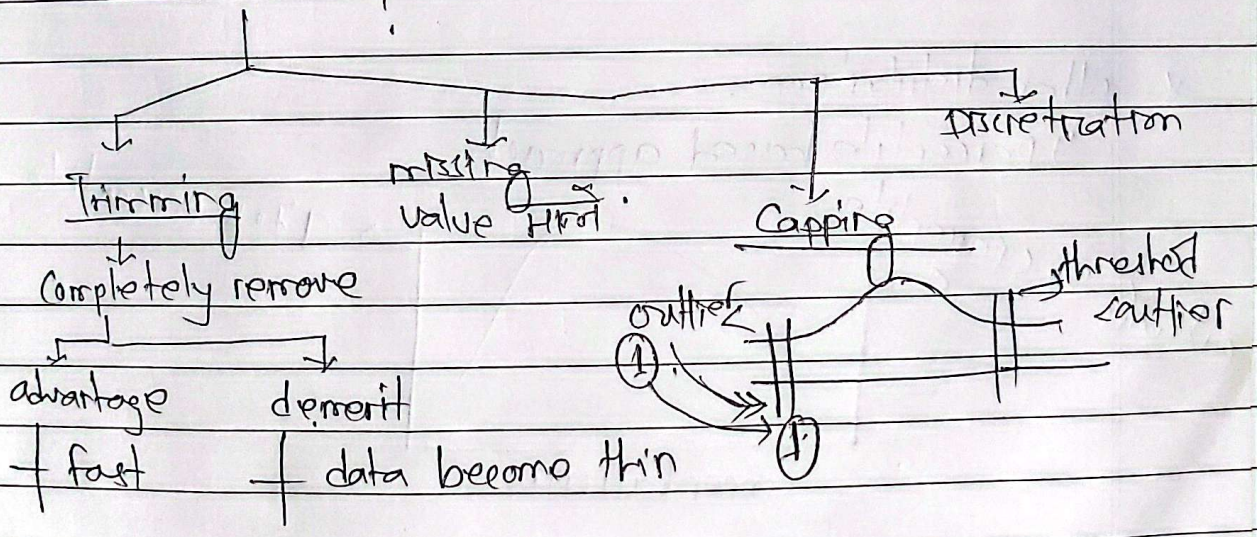
Effect of outliers in ML algorithms?

- Linear Regression
- logistic Regression
- Adaboost
- Deep learning.

} → weight based algorithm

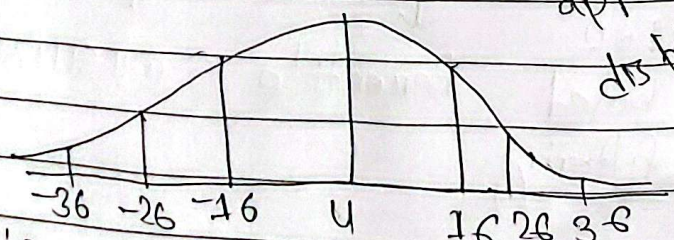
But dorot effect tree based algorithm like
decision tree
Random forest
Xgboost algorithm.

How to treat outliers?



How to detect outliers?

1. Normal distribution → but only applied on normally distributed data

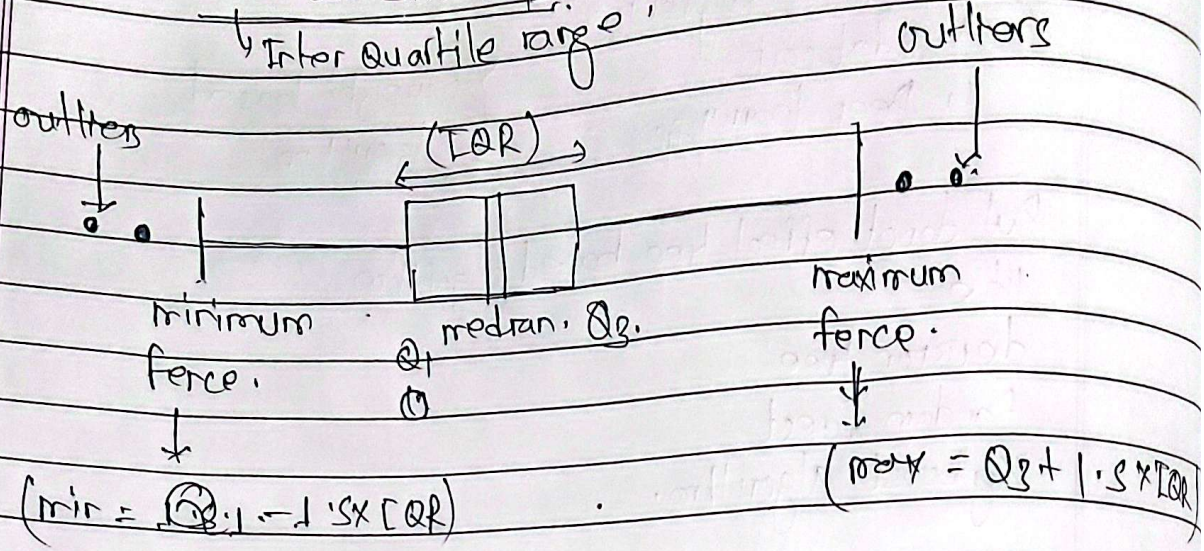


if ~~-3~~ (μ+3σ) < observation > (μ+3σ) → no outlier.
else outlier.

2) Skewed Distribution

↳ IQR use kiye |

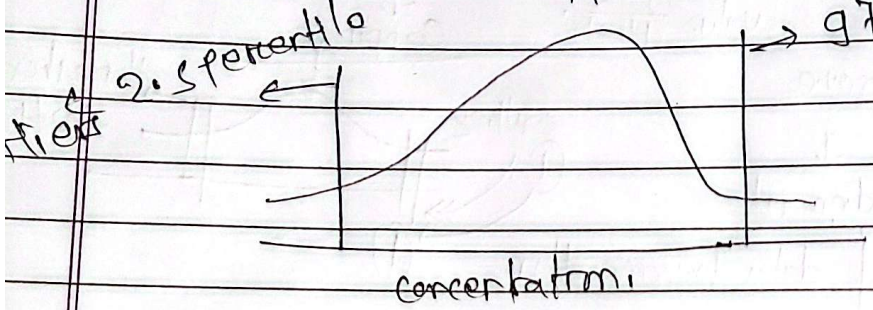
↳ Inter Quartile range.



only used when observation is skewed.

* other distribution

↳ percentile based approach

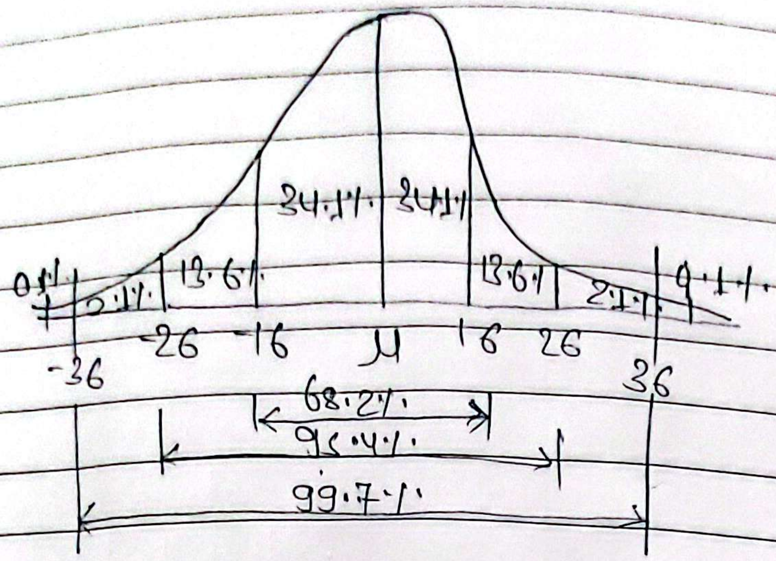


Sol

Techniques

- Z-score treatment
- IQR based filtering
- Percentile
- winsorization

outlier removal using Z-score.



$\left\{ \begin{matrix} \mu + \sigma \\ \mu - \sigma \end{matrix} \right\} \rightarrow$ 68% data.

$\left\{ \begin{matrix} \mu + 2\sigma \\ \mu - 2\sigma \end{matrix} \right\}$ 95% data.

So, outliers are value out of $\mu - 3\sigma$ & $\mu + 3\sigma$. or,

$-3 < \text{Zscore} < 3$ → range for outlier detection.

Z-Score?

$x_i' = \dfrac{x_i - \mu}{\sigma}$ for detection

outliers Treatment

- Trimming
  - ↳ dangerous (?)
  - ↳ significant data loss.
- capping

→ should be normally distributed or closed to normal distribution

* capping.

→ 85, 3, 90

80 5 80