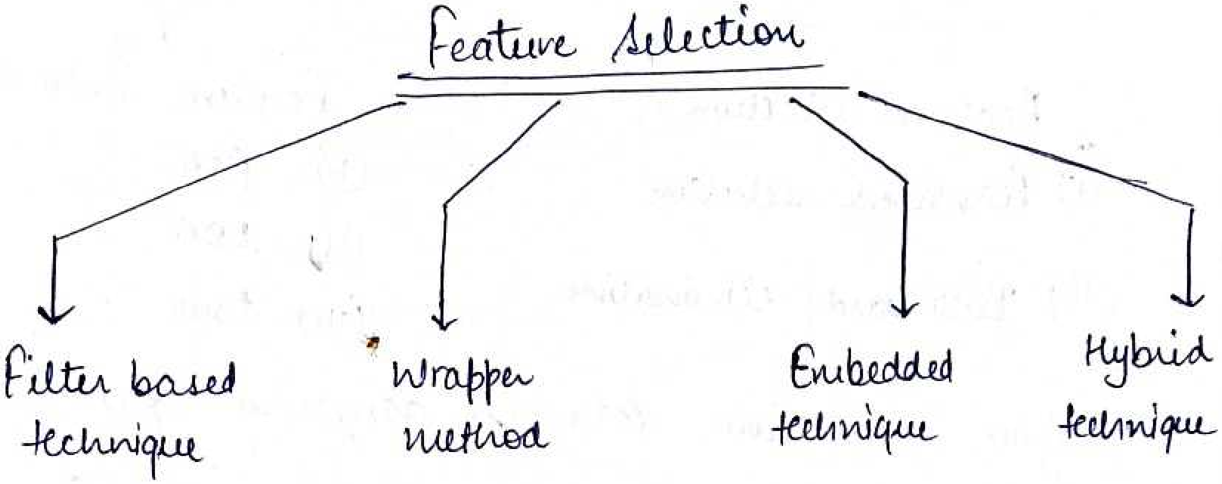
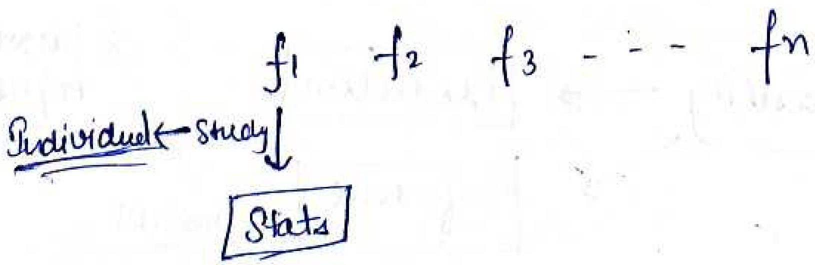
Types of Feature Selection



1.) **Filter based feature selection** :- Filter based feature selection techniques are methods that use statistical measure to score each feature independently, and then select a subset of features based on these scores. These method are called "filter" methods because they essentially filter out the features that do not meet some criterion.



Techniques

- Variance Threshold
- Correlation
- Anova
- Chi square
- Mutual info

1. Check Duplicate Features

f1	f2	f3	f4
1	2	1	2
2	1	2	2
3	3	3	2

Same.

dict → keep key → duplicate values

key → original cols

value → dublicate

2. Variance Threshold

→ Constant

→ quasi constant

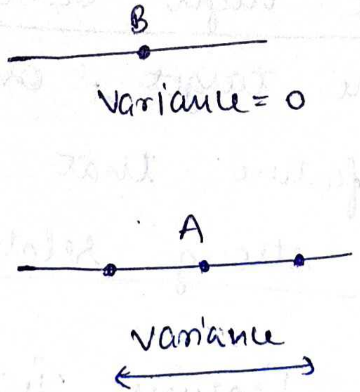
Constant

Drop

A	B	C
1	1	X
2	1	N
3	1	X

0 variance

constant always 1



quasi constant

Total 1000 rows

→ 995 rows = 1

→ 5 rows = 0 } variance ≃ 0 (quasi constant feature)

1) threshold = 0.1

2) fi → var (Check variance of every columns)

3) Drop those ~~variable~~ feature variance which is less than threshold.

low variance ← F1 ↳ relate to Y

f1 → Y

variance ← ↳ but not relate with Y
↓
high

Points to Consider

1. **Ignore Target Variable** :- Variance threshold is a individual univariate method, meaning it evalutates each feature independently and doesn't consider the relationship between each feature and the target variable. This means it may keep irrelevant features that have a high variance but no relationship with the target, or discard potentially useful feature that have a low variance but a strong relationship with target.

2. **Ignores Features Interactions**: Variance Threshold doesn't account for interaction between features. A features with a low variance may become very informative when combined with another feature.

f1 and f2 relation ↑↑

varian↓ ← f1 ↳ Variance↑

feature ↑↑

3) **Sensitive to Data Scaling :-** Variance Threshold is sensitive to the scale of the data. if features are not on the same scale, the variance will naturally be higher for features with larger values. Therefore, it is important to standardize the feature before applying variance threshold. 10000 ← f1 f2 → 0.2

minor change ↳ var↑

4) **Arbitrary threshold Value :-** It's up to the user to define what constitutes a " low variance. The threshold is not always easy to defined and the optimal value can vary between datasets.

0.1 or 0.01 or 0.2 is better