GUIDANCE NOTES

FEATURE SELECTIONS TECHNIQUES

1. What is Feature Selection?

- Feature Selection is process of choosing only the relevant features (variables/columns) in dataset that actually impacts the target variable (dependent variable) leaving out all other redundant and irrelevant features.
- In simple terms, Feature selection subsets the impactful independent variables in a dataset.
- For example, in titanic dataset, the passenger's travelled class, sex, age has more impact on the target variable (contains survived or not) than the passengers name, passenger id, time of onboarding.

2. Why Feature Selection?

The feature selection process helps to improve a model's performance in many ways:

- Improved Model Accuracy
- Less Computational time
- Simplified Model
- Reducing data storage

3. Feature Selection Techniques

There are many different techniques/methods used for feature selection. Some commonly used are:

a) Filter Method:

- This method makes use of correlation between each feature and the target (Correlation – analyses linear relationship between two variables and returns a negative or positive value).
- The features having high correlation are considered relevant.
  Example: Titanic dataset (survived or not)

Passenger class	Target (1 or 0)	0.56
sex		0.60
Passenger Id		-0.30
Age		-0.54

From the table, we can see that passenger class and sex has positive correlation with the target variable. Hence those can be considered relevant.

b) Wrapper Method:

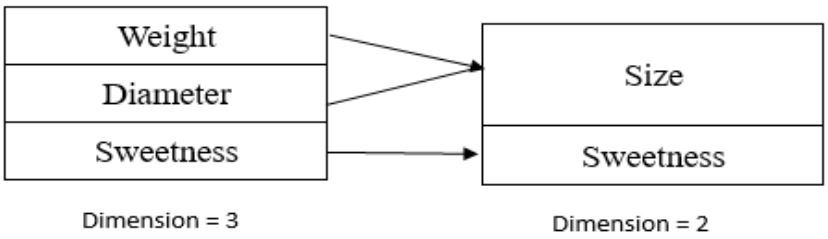
- This method trains the model repetitively with different set of features on each iteration and considers the combination of features that yield better performance as relevant features.
  Example: Titanic dataset (survived or not)

Features		Performance Score
1. Passenger Class 2. Passenger name	Model Training	0.69
1. Passenger Age 2. Passenger sex 3. Passenger class		0.86
1. Passenger Age 2. Passenger id 3. Passenger class		0.73

From the above table, the features subset with 0.86 gives better performance when compared to other subsets.

c) PCA:

- PCA is Principal Component Analysis which is dimension reduction technique.
- It simply combines two or more features with common or similar patterns and combines them into a new feature called the "Principal Component".
- We can also say that PCA summarizes the features into a new feature thus reducing dimension of the dataset.
  Example: Fruits dataset



The features "weight" and "diameter" have common variance; thus, those can be interpreted into a single feature called "Size" which contains both the features' patterns and data. (Here Size is the principal component)

4. Summary:

- Feature Selection is choosing features that are relevant and has impact on target variable.
- It helps in improving performance and data storage.
- Filter methods uses correlation to choose relevant features.
- Wrapper method uses different subset of features at each iteration and compares.
- PCA transforms original dataset into reduced dimensional representation.