Subset selection

Feature Selection Techniques in Machine Learning

- Feature selection is a way of selecting the subset of the most relevant features from the original features set by removing the redundant, irrelevant, or noisy features.
- While developing the machine learning model, only a few variables in the dataset are useful for building the model, and the rest features are either redundant or irrelevant.
- If we input the dataset with all these redundant and irrelevant features, it may negatively impact and reduce the overall performance and accuracy of the model.

- Hence it is very important to identify and select the most appropriate features from the data and remove the irrelevant or less important features, which is done with the help of feature selection in machine learning.
- Feature selection is one of the important concepts of machine learning, which highly impacts the performance of the model.
- As machine learning works on the concept of "Garbage In Garbage Out", so we always need to input the most appropriate and relevant dataset to the model in order to get a better

But before that, let's first understand some basics of feature selection.

- What is Feature Selection?
- Need for Feature Selection
- Feature Selection Methods/Techniques
- Feature Selection statistics

What is Feature Selection?

- A feature is an attribute that has an impact on a problem or is useful for the problem, and choosing the important features for the model is known as feature selection.
- Each machine learning process depends on feature engineering, which mainly contains two processes; which are Feature Selection and Feature Extraction.
- Although feature selection and extraction processes may have the same objective, both are completely different from each other.

- The main difference between them is that feature selection is about selecting the subset of the original feature set, whereas feature extraction creates new features.

Need for Feature Selection

- Before implementing any technique, it is really important to understand, need for the technique and so for the Feature Selection.
- As we know, in machine learning, it is necessary to provide a pre-processed and good input dataset in order to get better outcomes.
- We collect a huge amount of data to train our model and help it to learn better.
- Generally, the dataset consists of noisy data, irrelevant data, and some part of useful data.
- Moreover, the huge amount of data also slows down the training process of the model, and with noise and irrelevant data, the model may not predict and perform well.
- So, it is very necessary to remove such noises and less-important data from the dataset and to do this, and Feature selection techniques are used.

- Selecting the best features helps the model to perform well.

For example:

- Suppose we want to create a model that automatically decides which car should be crushed for a spare part, and to do this, we have a dataset.
- This dataset contains a Model of the car, Year, Owner's name, Miles.
- So, in this dataset, the name of the owner does not contribute to the model performance as it does not decide if the car should be crushed or not, so we can remove this column and select the rest of the features(column) for the model building.

Below are some benefits of using feature selection in machine learning:

- It helps in avoiding the curse of dimensionality.

- It helps in the simplification of the model so that it can be easily interpreted by the researchers.
- It reduces the training time.
- It reduces overfitting hence enhance the generalization.

Feature Selection Techniques

- There are mainly two types of Feature Selection techniques, which are:

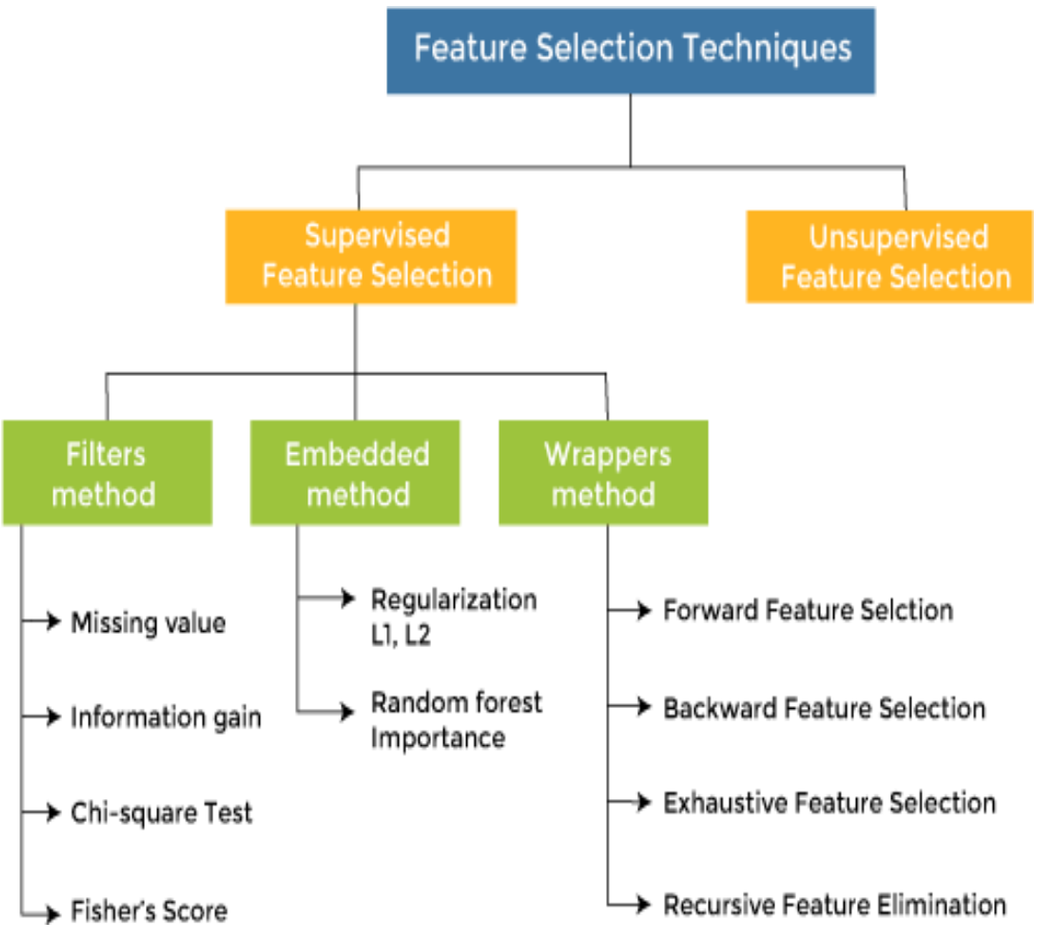
- **Supervised Feature Selection technique**

  Supervised Feature selection techniques consider the target variable and can be used for the labelled dataset.

- **Unsupervised Feature Selection technique**

  Unsupervised Feature selection techniques ignore the target variable and can be used for the unlabelled

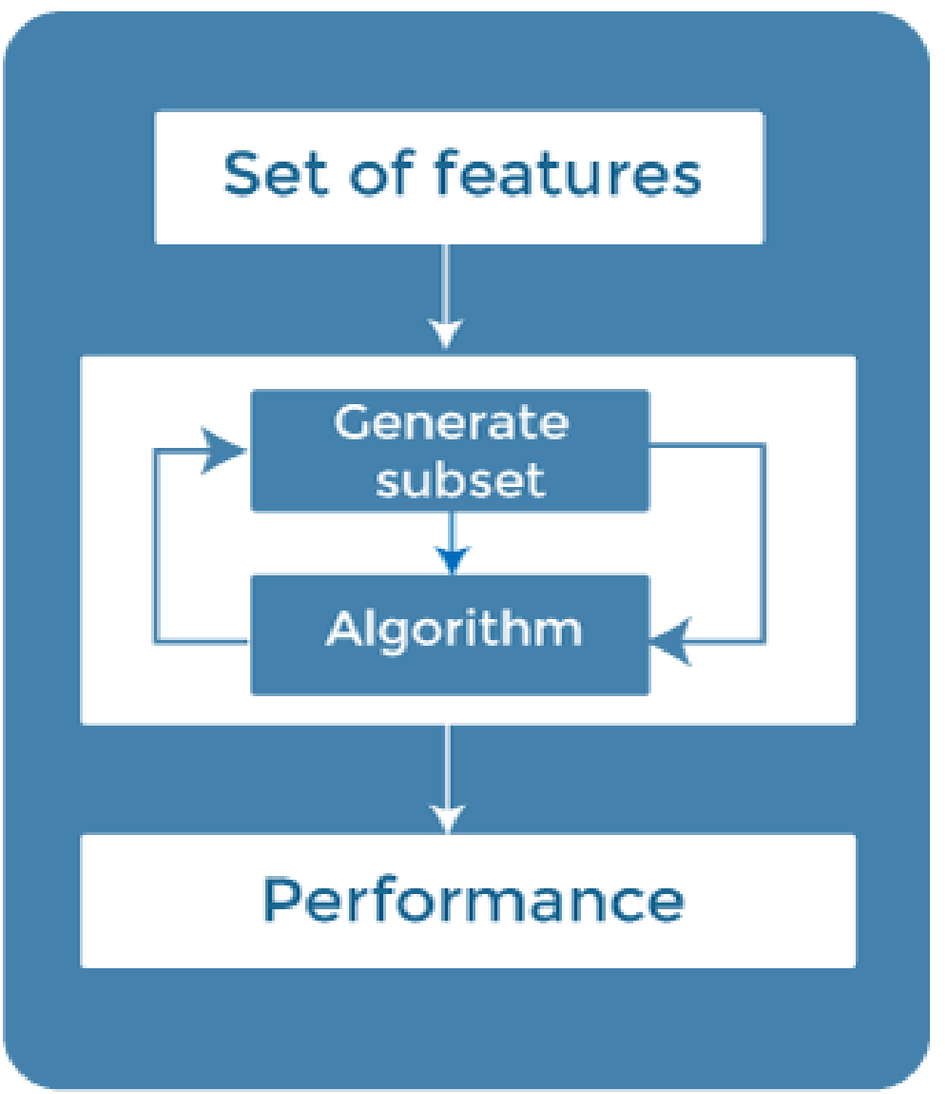
dataset.



1. Wrapper Methods

- In wrapper methodology, selection of features is done by considering it as a search problem, in which different combinations are made, evaluated, and compared with other combinations.
- It trains the algorithm by using the subset of features iteratively.

- On the basis of the output of the model, features are added or subtracted, and with this feature set, the model has trained again.



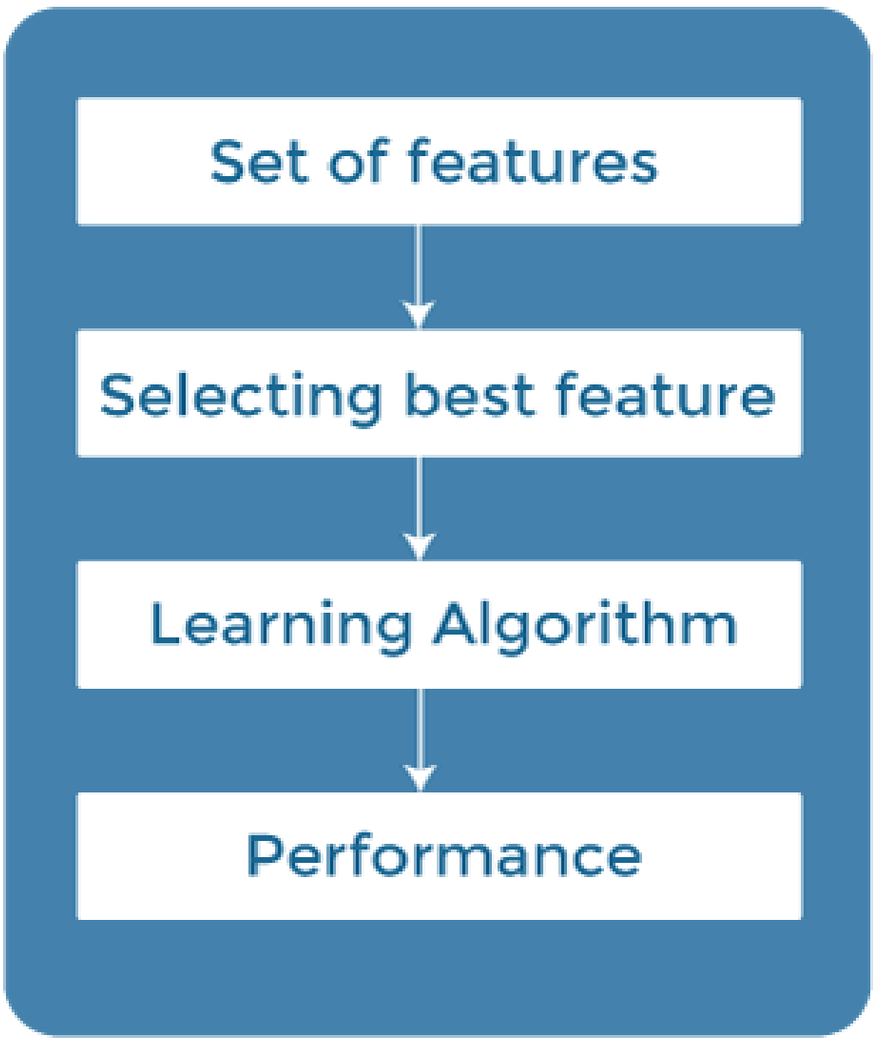
Some techniques of wrapper methods are:

- **Forward selection**

- **Backward elimination**
- **Exhaustive Feature Selection**
- **Recursive Feature Elimination**

2. Filter Methods

- In Filter Method, features are selected on the basis of statistics measures.
- This method does not depend on the learning algorithm and chooses the features as a pre- processing step.
- The filter method filters out the irrelevant feature and redundant columns from the model by using different metrics through ranking.
- The advantage of using filter methods is that it needs low computational time and does not overfit the data.



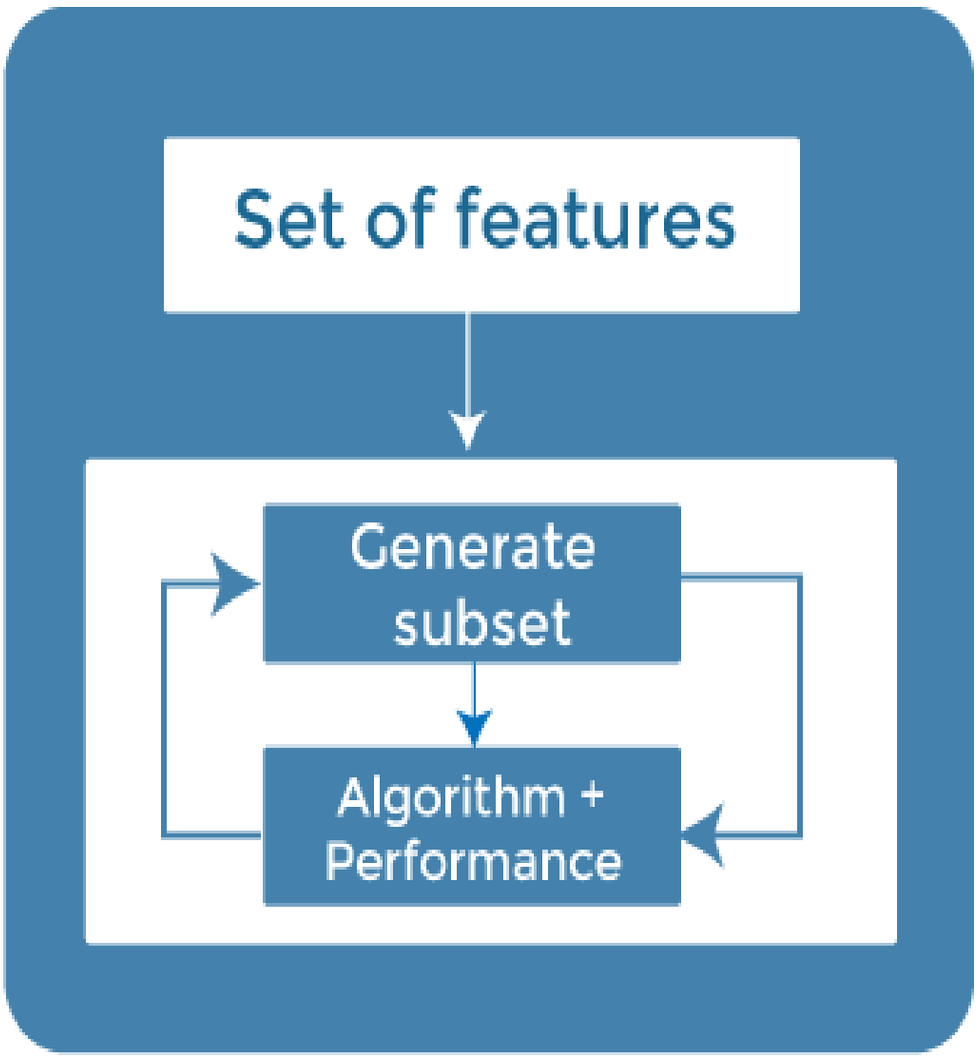
Some common techniques of Filter methods are as follows:

- Information Gain
- Chi-square Test
- Fisher's Score

- .

3. Embedded Methods

- Embedded methods combined the advantages of both filter and wrapper methods by considering the interaction of features along with low computational cost.
- These are fast processing methods similar to the filter method but more accurate than the filter method.
- These methods are also iterative, which evaluates each iteration, and optimally finds the most important features that contribute the most to training in a particular iteration.


Set of features
Generate subset
Algorithm + Performance

Some techniques of embedded methods are:

- Regularization
- Random Forest Importance

- Feature selection is a very complicated and vast field of machine learning, and lots of studies are already made to discover the best methods.
- There is no fixed rule of the best feature selection method. However, choosing the method depend on a machine learning engineer who can combine and innovate approaches to find the best method for a specific problem.
- One should try a variety of model fits on different subsets of features selected through different statistical Measures.