Feature Selection :-

- So, we can define feature Selection as, "It is a process of automatically or manually selecting the subset of most appropriate and relevant features to be used in model building." Feature selection is performed by either including the important features or excluding the irrelevant features in the dataset without changing them.

Need for Feature Selection

- 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:-

1. wrapper method

a.Forward Selection:-
- Forward selection is an iterative method in which we start with having no feature in the model. In each iteration, we keep adding the feature which best improves our model till an addition of a new variable does not improve the performance of the model.
- Forward selection is an iterative process, which begins with an empty set of features. After each iteration, it keeps adding on a feature and evaluates the performance to check whether it is improving the performance or not. The process continues until the addition of a new variable/feature does not improve the performance of the model.
b.Backward Elimination:-
- In backward elimination, we start with all the features and removes the least significant feature at each iteration which improves the performance of the model. We repeat this until no improvement is observed on removal of features.
- all festure give to the model - statitical test - chi-square test- determine the p-value - if p value is less than 0.5 the feature is useful - if the p value is greater than 0.5 then the feature is not useful or irelevent we can remove it.
c.Exhaustive Feature Selection :-
- Exhaustive feature selection is one of the best feature selection methods, which evaluates each feature set as brute-force. It means this method tries & make each possible combination of features and return the best performing feature set.
d.Recursive Feature Elimination-
- Recursive feature elimination is a recursive greedy optimization approach, where features are selected by recursively taking a smaller and smaller subset of features. Now, an estimator is trained with each set of features, and the importance of each feature is determined using coef_attribute or through a feature_importances_attribute.

2. Filter Methods :-

- 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.

a.Information Gain:

- Information gain determines the reduction in entropy while transforming the dataset. It can be used as a feature selection technique by calculating the information gain of each variable with respect to the target variable.

b.Chi-square Test:

- Chi-square test is a technique to determine the relationship between the categorical variables. The chi-square value is calculated between each feature and the target variable, and the desired number of features with the best chi-square value is selected.

c.Fisher's Score:

- Fisher's score is one of the popular supervised technique of features selection. It returns the rank of the variable on the fisher's criteria in descending order. Then we can select the variables with a large fisher's score.

d.Missing Value Ratio:

- The value of the missing value ratio can be used for evaluating the feature set against the threshold value. The formula for obtaining the missing value ratio is the number of missing values in each column divided by the total number of observations. The variable is having more than the threshold value can be dropped.

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.

```
  a.Regularization-
   - Regularization adds a penalty term to different parameters of the m
achine learning model for avoiding overfitting in the model. This penalt
y term is added to the coefficients; hence it shrinks some coefficients
 to zero. Those features with zero coefficients can be removed from the
 dataset. The types of regularization techniques are L1 Regularization
 (Lasso Regularization) or Elastic Nets (L1 and L2 regularization).

  b.Random Forest Importance :-
  - Different tree-based methods of feature selection help us with featu
re importance to provide a way of selecting features. Here, feature impo
rtance specifies which feature has more importance in model building or
 has a great impact on the target variable. Random Forest is such a tree
-based method, which is a type of bagging algorithm that aggregates a di
fferent number of decision trees. It automatically ranks the nodes by th
eir performance or decrease in the impurity (Gini impurity) over all the
trees. Nodes are arranged as per the impurity values, and thus it allows
to pruning of trees below a specific node. The remaining nodes create a
 subset of the most important features.
```

How to choose a Feature Selection Method?

- Numerical Variables: Variable with continuous values such as integer, float
- Categorical Variables: Variables with categorical values such as Boolean, ordinal, nominals.

1. Numerical Input, Numerical Output:

- Numerical Input variables are used for predictive regression modelling. The common method to be used for such a case is the Correlation coefficient.
  - Pearson's correlation coefficient (For linear Correlation).
    - Spearman's rank coefficient (for non-linear correlation).

2. Numerical Input, Categorical Output:

- Numerical Input with categorical output is the case for classification predictive modelling problems. In this case, also, correlation-based techniques should be used, but with categorical output.
  - ANOVA correlation coefficient (linear).
  - Kendall's rank coefficient (nonlinear).

3. Categorical Input, Numerical Output:

- This is the case of regression predictive modelling with categorical input. It is a different example of a regression problem. We can use the same measures as discussed in the above case but in reverse order.
  - Kendall's rank coefficient (linear).
  - ANOVA correlation coefficient (nonlinear).

4. Categorical Input, Categorical Output:

- This is a case of classification predictive modelling with categorical Input variables.
  - Chi-Squared test (contingency tables).
  - Mutual Information.

In [ ]:

In [ ]: