**Feature selection:**

Feature selection is a process that chooses a subset of features from the original features so that the feature space is optimally reduced according to a certain criterion.

**Feature selection** methods are intended to reduce the number of input variables to those that are believed to be most useful to a model in order to predict the target variable.

*Feature selection is primarily focused on removing non-informative or redundant predictors from the model.*

Some predictive modeling problems have a large number of variables that can slow the development and training of models and require a large amount of system memory. Additionally, the performance of some models can degrade when including input variables that are not relevant to the target variable.

*Many models, especially those based on regression slopes and intercepts, will estimate parameters for every term in the model. Because of this, the presence of non-informative variables can add uncertainty to the predictions and reduce the overall effectiveness of the model.*

One way to think about feature selection methods are in terms of **supervised** and **unsupervised** methods.

*An important distinction to be made in feature selection is that of supervised and unsupervised methods. When the outcome is ignored during the elimination of predictors, the technique is unsupervised.*

The difference has to do with whether features are selected based on the target variable or not. Unsupervised feature selection techniques ignores the target variable, such as methods that remove redundant variables using correlation. Supervised feature selection techniques use the target variable, such as methods that remove irrelevant variables..

Another way to consider the mechanism used to select features which may be divided into **wrapper** and **filter** methods. These methods are almost always supervised and are evaluated based on the performance of a resulting model on a hold out dataset.

*Filter methods evaluate the relevance of the predictors outside of the predictive models and subsequently model only the predictors that pass some criterion.*

Finally, there are some machine learning algorithms that perform feature selection automatically as part of learning the model. We might refer to these techniques as **intrinsic** feature selection methods.

*… some models contain built-in feature selection, meaning that the model will only include predictors that help maximize accuracy. In these cases, the model can pick and choose which representation of the data is best.*

This includes algorithms such as penalized regression models like Lasso and decision trees, including ensembles of decision trees like random forest.

*Some models are naturally resistant to non-informative predictors. Tree- and rule-based models, MARS and the lasso, for example, intrinsically conduct feature selection.*

Feature selection is also related to [dimensionally reduction](https://machinelearningmastery.com/dimensionality-reduction-for-machine-learning/) techniques in that both methods seek fewer input variables to a predictive model. The difference is that feature selection select features to keep or remove from the dataset, whereas dimensionality reduction create a projection of the data resulting in entirely new input features. As such, dimensionality reduction is an alternate to feature selection rather than a type of feature selection.

 Feature selection is a critical step in the feature construction process

There are three general classes of feature selection algorithms: **Filter methods, wrapper methods and embedded methods**.

The role of feature selection in machine learning is,

1. To reduce the dimensionality of feature space.

2. To speed up a learning algorithm.

3. To improve the predictive accuracy of a classification algorithm.

4. To improve the comprehensibility of the learning results.

**Features Selection Algorithms are as follows:**

**1**. **Instance based approaches:** There is no explicit procedure for feature subset generation. Many small data samples are sampled from the data. Features are weighted according to their roles in differentiating instances of different classes for a data sample. Features with higher weights can be selected.

**2. Nondeterministic approaches:**Genetic algorithms and simulated annealing are also used in feature selection.

**3. Exhaustive complete approaches:**Branch and Bound evaluates estimated accuracy and ABB checks an inconsistency measure that is monotonic. Both start with a full feature set until the preset bound cannot be maintained.

While building a machine learning model for real-life dataset, we come across a lot of features in the dataset and not all these features are important every time. Adding unnecessary features while training the model leads us to reduce the overall accuracy of the model, increase the complexity of the model and decrease the generalization capability of the model and makes the model biased. Even the saying “Sometimes less is better” goes as well for the machine learning model. Hence, **feature selection** is one of the important steps while building a machine learning model. Its goal is to find the best possible set of features for building a machine learning model.

Some popular techniques of feature selection in machine learning are:

* Filter methods
* Wrapper methods
* Embedded methods

**Filter Methods**

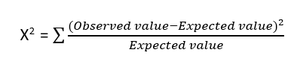
These methods are generally used while doing the pre-processing step. These methods select features from the dataset irrespective of the use of any machine learning algorithm. In terms of computation, they are very fast and inexpensive and are very good for removing duplicated, correlated, redundant features but these methods do not remove multicollinearity. Selection of feature is evaluated individually which can sometimes help when features are in isolation (don’t have a dependency on other features) but will lag when a combination of features can lead to increase in the overall performance of the model.



*Filter Methods Implementation*

Some techniques used are:

* **Information Gain –** It is defined as the amount of information provided by the feature for identifying the target value and measures reduction in the entropy values. Information gain of each attribute is calculated considering the target values for feature selection.
* **Chi-square test —** Chi-square method (X2) is generally used to test the relationship between categorical variables. It compares the observed values from different attributes of the dataset to its expected value.



*Chi-square Formula*

* **Fisher’s Score –** Fisher’s Score selects each feature independently according to their scores under Fisher criterion leading to a suboptimal set of features. The larger the Fisher’s score is, the better is the selected feature.
* **Correlation Coefficient –** Pearson’s Correlation Coefficient is a measure of quantifying the association between the two continuous variables and the direction of the relationship with its values ranging from *-1 to 1*.
* **Variance Threshold –** It is an approach where all features are removed whose variance doesn’t meet the specific threshold. By default, this method removes features having zero variance. The assumption made using this method is higher variance features are likely to contain more information.
* **Mean Absolute Difference (MAD) –** This method is similar to variance threshold method but the difference is there is no square in MAD. This method calculates the mean absolute difference from the mean value.
* **Dispersion Ratio –** Dispersion ratio is defined as the ratio of the Arithmetic mean (AM) to that of Geometric mean (GM) for a given feature. Its value ranges from *+1 to ∞ as AM ≥ GM* for a given feature. Higher dispersion ratio implies a more relevant feature.
* **Mutual Dependence –**This method measures if two variables are mutually dependent, and thus provides the amount of information obtained for one variable on observing the other variable. Depending on the presence/absence of a feature, it measures the amount of information that feature contributes to making the target prediction.
* **Relief –** This method measures the quality of attributes by randomly sampling an instance from the dataset and updating each feature and distinguishing between instances that are near to each other based on the difference between the selected instance and two nearest instances of same and opposite classes.

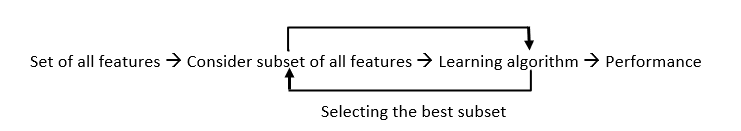
**Wrapper methods:**

Wrapper methods, also referred to as greedy algorithms train the algorithm by using a subset of features in an iterative manner. Based on the conclusions made from training in prior to the model, addition and removal of features takes place. Stopping criteria for selecting the best subset are usually pre-defined by the person training the model such as when the performance of the model decreases or a specific number of features has been achieved. The main advantage of wrapper methods over the filter methods is that they provide an optimal set of features for training the model, thus resulting in better accuracy than the filter methods but are computationally more expensive.

Wrapper feature selection methods create many models with different subsets of input features and select those features that result in the best performing model according to a performance metric. These methods are unconcerned with the variable types, although they can be computationally expensive. [RFE](https://scikit-learn.org/stable/modules/generated/sklearn.feature_selection.RFE.html) is a good example of a wrapper feature selection method.

*Wrapper methods evaluate multiple models using procedures that add and/or remove predictors to find the optimal combination that maximizes model performance.*

Filter feature selection methods use statistical techniques to evaluate the relationship between each input variable and the target variable, and these scores are used as the basis to choose (filter) those input variables that will be used in the model.



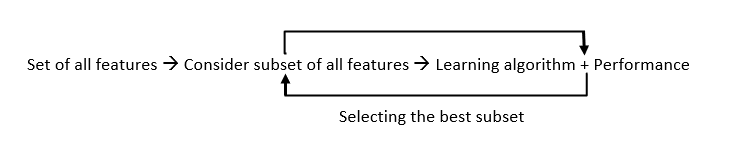
*Wrapper Methods Implementation*

Some techniques used are:

* **Forward selection –**This method is an iterative approach where we initially start with an empty set of features and keep adding a feature which best improves our model after each iteration. The stopping criterion is till the addition of a new variable does not improve the performance of the model.
* **Backward elimination –** This method is also an iterative approach where we initially start with all features and after each iteration, we remove the least significant feature. The stopping criterion is till no improvement in the performance of the model is observed after the feature is removed.
* **Bi-directional elimination –** This method uses both forward selection and backward elimination technique simultaneously to reach one unique solution.
* **Exhaustive selection –** This technique is considered as the brute force approach for the evaluation of feature subsets. It creates all possible subsets and builds a learning algorithm for each subset and selects the subset whose model’s performance is best.
* **Recursive elimination –** This greedy optimization method selects features by recursively considering the smaller and smaller set of features. The estimator is trained on an initial set of features and their importance is obtained using feature\_importance\_attribute. The least important features are then removed from the current set of features till we are left with the required number of features.

**Embedded methods:**

In embedded methods, the feature selection algorithm is blended as part of the learning algorithm, thus having its own built-in feature selection methods. Embedded methods encounter the drawbacks of filter and wrapper methods and merge their advantages. These methods are faster like those of filter methods and more accurate than the filter methods and take into consideration a combination of features as well.



*Embedded Methods Implementation*

Some techniques used are:

* **Regularization –** This method adds a penalty to different parameters of the machine learning model to avoid over-fitting of the model. This approach of feature selection uses Lasso (L1 regularization) and Elastic nets (L1 and L2 regularization). The penalty is applied over the coefficients, thus bringing down some coefficients to zero. The features having zero coefficient can be removed from the dataset.
* **Tree-based methods –**These methods such as Random Forest, Gradient Boosting provides us feature importance as a way to select features as well. Feature importance tells us which features are more important in making an impact on the target feature.

**Conclusion:**

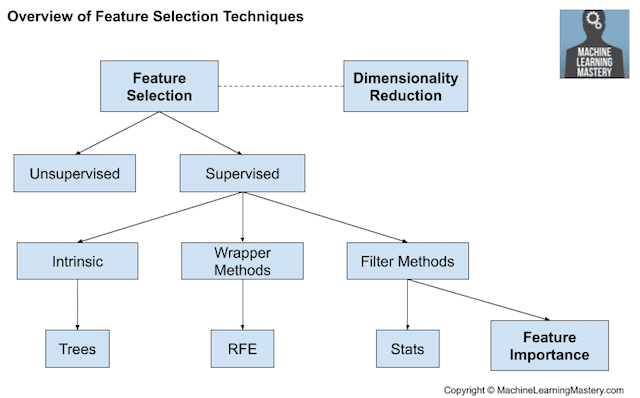
Apart from the methods discussed above, there are many other methods of feature selection. Using hybrid methods for feature selection can offer a selection of best advantages from other methods, leading to reduce in the disadvantages of the algorithms. These models can provide greater accuracy and performance when compared to other methods. Dimensionality reduction techniques such as Principal Component Analysis (PCA), Heuristic Search Algorithms, etc. don’t work in the way as to feature selection techniques but can help us to reduce the number of features.

Feature selection is a wide, complicated field and a lot of studies has already been made to figure out the best methods. It depends on the machine learning engineer to combine and innovate approaches, test them and then see what works best for the given problem.

We can summarize feature selection as follows.

* **Feature Selection**: Select a subset of input features from the dataset.
  + **Unsupervised**: Do not use the target variable (e.g. remove redundant variables).
    - Correlation
  + **Supervised**: Use the target variable (e.g. remove irrelevant variables).
    - **Wrapper**: Search for well-performing subsets of features.
      * RFE
    - **Filter**: Select subsets of features based on their relationship with the target.
      * Statistical Methods
      * Feature Importance Methods
    - **Intrinsic**: Algorithms that perform automatic feature selection during training.
      * Decision Trees
* **Dimensionality Reduction**: Project input data into a lower-dimensional feature space.

The image below provides a summary of this hierarchy of feature selection techniques.



Overview of Feature Selection Techniques

**2. Statistics for Filter-Based Feature Selection Methods**

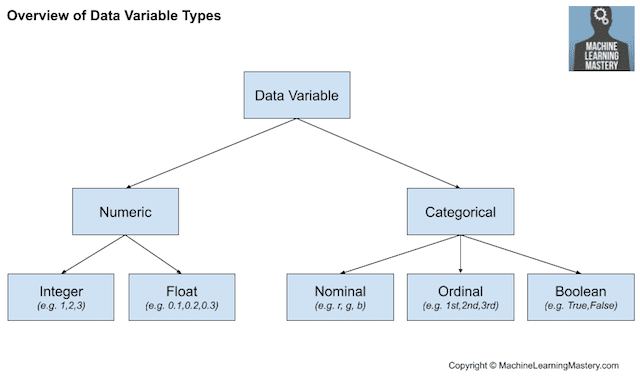
It is common to use correlation type statistical measures between input and output variables as the basis for filter feature selection.

As such, the choice of statistical measures is highly dependent upon the variable data types.

Common data types include numerical (such as height) and categorical (such as a label), although each may be further subdivided such as integer and floating point for numerical variables, and boolean, ordinal, or nominal for categorical variables.

Common input variable data types:

* **Numerical Variables**
  + Integer Variables.
  + Floating Point Variables.
* **Categorical Variables**.
  + Boolean Variables (dichotomous).
  + Ordinal Variables.
  + Nominal Variables.



The more that is known about the data type of a variable, the easier it is to choose an appropriate statistical measure for a filter-based feature selection method.

In this section, we will consider two broad categories of variable types: numerical and categorical; also, the two main groups of variables to consider: input and output.

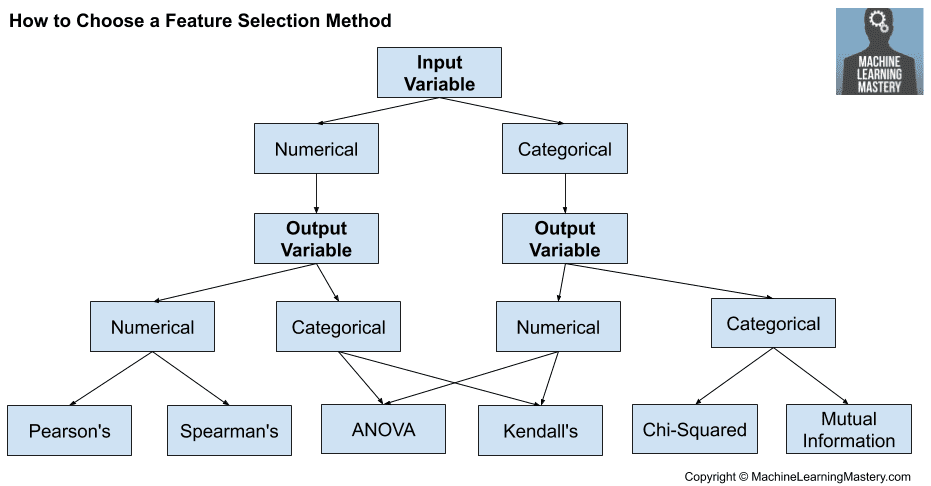
Input variables are those that are provided as input to a model. In feature selection, it is this group of variables that we wish to reduce in size. Output variables are those for which a model is intended to predict, often called the response variable.

The type of response variable typically indicates the type of predictive modeling problem being performed. For example, a numerical output variable indicates a regression predictive modeling problem, and a categorical output variable indicates a classification predictive modeling problem.

* **Numerical Output**: Regression predictive modeling problem.
* **Categorical Output**: Classification predictive modeling problem.

The statistical measures used in filter-based feature selection are generally calculated one input variable at a time with the target variable. As such, they are referred to as univariate statistical measures. This may mean that any interaction between input variables is not considered in the filtering process.

*Most of these techniques are univariate, meaning that they evaluate each predictor in isolation. In this case, the existence of correlated predictors makes it possible to select important, but redundant, predictors. The obvious consequences of this issue are that too many predictors are chosen and, as a result, collinearity problems arise.*



**Numerical Input, Numerical Output**

This is a regression predictive modeling problem with numerical input variables.

The most common techniques are to use a correlation coefficient, such as Pearson’s for a linear correlation, or rank-based methods for a nonlinear correlation.

* Pearson’s correlation coefficient (linear).
* Spearman’s rank coefficient (nonlinear)

**Numerical Input, Categorical Output**

This is a classification predictive modeling problem with numerical input variables.

This might be the most common example of a classification problem,

Again, the most common techniques are correlation based, although in this case, they must take the categorical target into account.

* ANOVA correlation coefficient (linear).
* Kendall’s rank coefficient (nonlinear).

Kendall does assume that the categorical variable is ordinal.

**Categorical Input, Numerical Output**

This is a regression predictive modeling problem with categorical input variables.

This is a strange example of a regression problem (e.g. you would not encounter it often).

Nevertheless, you can use the same “*Numerical Input, Categorical Output*” methods (described above), but in reverse.

**Categorical Input, Categorical Output**

This is a classification predictive modeling problem with categorical input variables.

The most common correlation measure for categorical data is the [chi-squared test](https://machinelearningmastery.com/chi-squared-test-for-machine-learning/). You can also use mutual information (information gain) from the field of information theory.

* Chi-Squared test (contingency tables).
* Mutual Information.

In fact, mutual information is a powerful method that may prove useful for both categorical and numerical data, e.g. it is agnostic to the data types.

[How to Choose a Feature Selection Method For Machine Learning - MachineLearningMastery.com](https://machinelearningmastery.com/feature-selection-with-real-and-categorical-data/)

[Feature Selection Techniques in Machine Learning with Python | by Rahil Shaikh | Towards Data Science](https://towardsdatascience.com/feature-selection-techniques-in-machine-learning-with-python-f24e7da3f36e)

[An Introduction to Feature Selection - MachineLearningMastery.com](https://machinelearningmastery.com/an-introduction-to-feature-selection/)

[Feature selection techniques for classification and Python tips for their application | by Gabriel Azevedo | Towards Data Science](https://towardsdatascience.com/feature-selection-techniques-for-classification-and-python-tips-for-their-application-10c0ddd7918b)

[2106.06437 (arxiv.org)](https://arxiv.org/pdf/2106.06437)