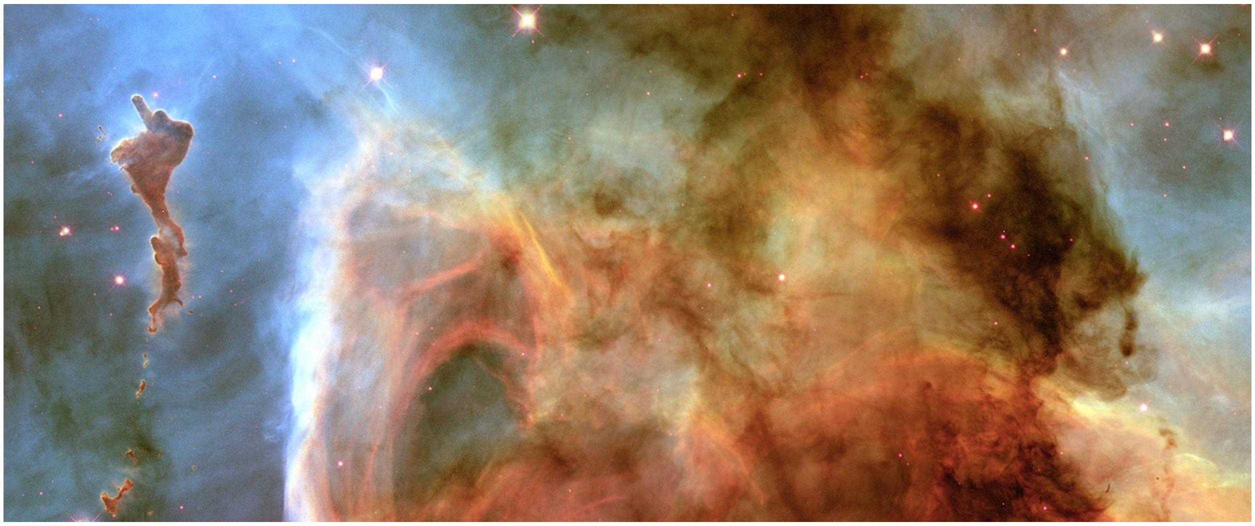


Ext Feature selection :

Some techniques used are:

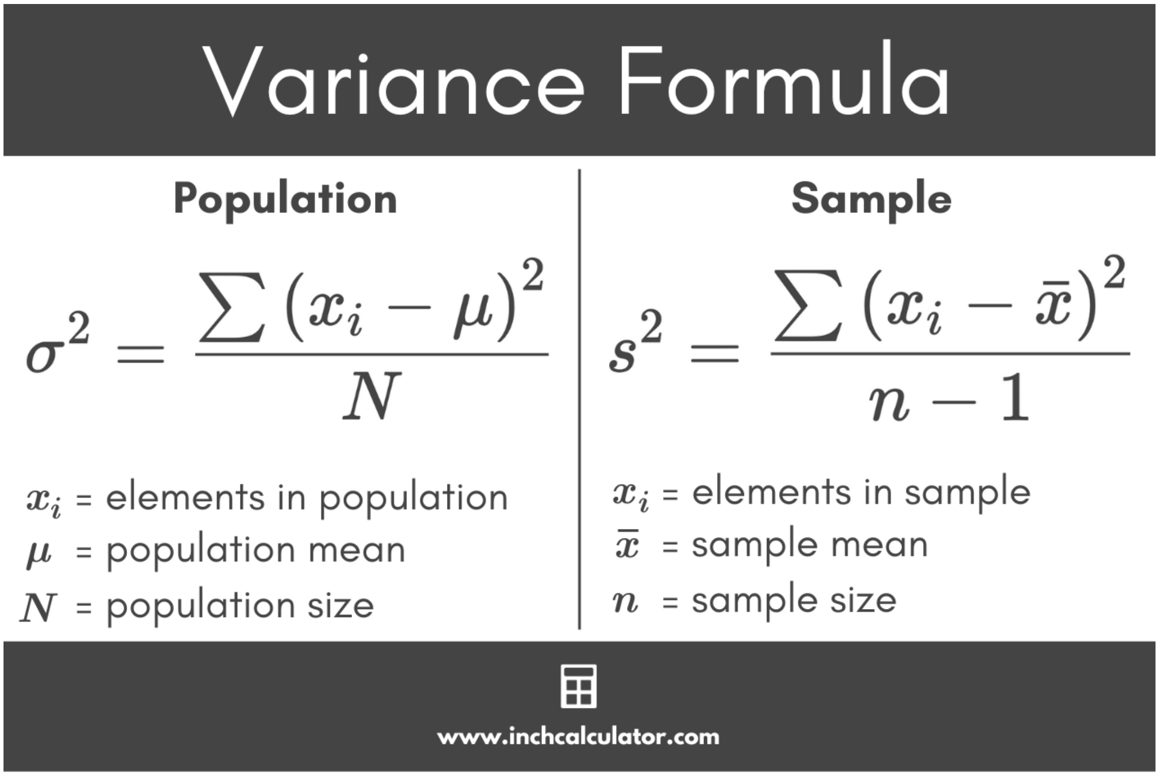
- **Information Gain** **–** It is defined as the amount of information provided by the feature for identifying the target value and measures reduction in the entropy values. Information gain of each attribute is calculated considering the target values for feature selection.
- **Chi-square test** **—** 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.

$$X^2 = \sum \frac{(Observed\ value - Expected\ value)^2}{Expected\ value}$$

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.
- Reminder :



Variance Formula

Population

$$\sigma^2 = \frac{\sum (x_i - \mu)^2}{N}$$

x_i = elements in population

μ = population mean

N = population size

Sample

$$s^2 = \frac{\sum (x_i - \bar{x})^2}{n - 1}$$

x_i = elements in sample

$\bar{x}$ = sample mean

n = sample size

www.inchcalculator.com

SelectKBest:

SelectKBest is a feature selection technique used to select the top k features based on their scores.
It scores features using univariate statistical tests such as chi-squared test (for classification tasks), F-test (for regression tasks), or mutual information.
SelectKBest can be used for both classification and regression tasks.

chi2:

chi2 is a specific scoring function used in SelectKBest for feature selection.
It computes the chi-squared statistic and p-values between each feature and the target variable.
It is typically used for classification tasks to select features based on their independence with the target variable.

f_classif:

f_classif is another scoring function used in SelectKBest for feature selection.
It computes the ANOVA F-value between each feature and the target variable.

It is primarily used for classification tasks to select features based on their relationship with the target variable.

mutual_info_classif:

mutual_info_classif is a scoring function used in SelectKBest for feature selection. It computes the mutual information between each feature and the target variable. It is used for both classification and regression tasks to select features based on their information gain with respect to the target variable.

Usage:

SelectKBest: Can be used for both classification and regression tasks.
chi2: Typically used for classification tasks.
f_classif: Typically used for classification tasks.
mutual_info_classif: Can be used for both classification and regression tasks, but particularly useful for tasks where feature interactions need to be considered.