The train-test split procedure is used to estimate the performance of machine learning algorithms when they are used to make predictions on data not used to train the model.

It is a fast and easy procedure to perform, the results of which allow you to compare the performance of machine learning algorithms for your predictive modeling problem. Although simple to use and interpret, there are times when the procedure should not be used, such as when you have a small dataset and situations where additional configuration is required, such as when it is used for classification and the dataset is not balanced.

Ref: <https://machinelearningmastery.com/train-test-split-for-evaluating-machine-learning-algorithms/>

Another situation could be where you are not interested in predicting "new data", rather, just getting a feel for the relationships/patterns between the predictors and the response variable.

Ref: <https://stats.stackexchange.com/questions/290808/when-to-not-split-up-your-data-into-training-and-testing>

**Bayesian: a useful mechanism to deal with insufficient data, or poor distributed data;** to perform regression on a vector of basis functions

**https://stackoverflow.com/questions/13279213/compare-bayesian-linear-regression-vs-linear-regression**

If the dataset contains features some of which are Categorical Variables and some of the others are **continuous** variable Decision Tree is better than Linear Regression,since Trees can accurately divide the data based on Categorical Variables

Deep forest is lso used when Heterogeneous data are **any data with high variability of data types and formats**. They are possibly ambiguous and low quality due to missing values, high data redundancy, and untruthfulness.

<https://analyticsindiamag.com/a-beginners-guide-to-deep-regression-forest/>

random forest is is more good in in datasets for a very large datasetd. There is high dimensional data.

https://www.statology.org/high-dimensional-data/

However, they pose a major challenge that is that they can’t extrapolate outside unseen data

<https://neptune.ai/blog/random-forest-regression-when-does-it-fail-and-why>

<https://docs.microsoft.com/en-us/azure/machine-learning/component-reference/decision-forest-regression>

**For Our Model, is it necessary to put “Train\_Test\_Splits”?**

Graphical user interface, text, application

Description automatically generated

OLS Regression Results

==============================================================================

Dep. Variable: life\_ladder R-squared: 0.745

Model: OLS Adj. R-squared: 0.720

Method: Least Squares F-statistic: 29.88

Date: Thu, 27 Jan 2022 Prob (F-statistic): 2.80e-21

Time: 08:43:52 Log-Likelihood: -71.698

No. Observations: 91 AIC: 161.4

Df Residuals: 82 BIC: 184.0

Df Model: 8

Covariance Type: nonrobust

====================================================================================================

coef std err t P>|t| [0.025 0.975]

----------------------------------------------------------------------------------------------------

const 3.6031 0.804 4.479 0.000 2.003 5.203

democracy\_index 0.0066 0.005 1.474 0.144 -0.002 0.016

gender\_ratio\_males\_per100\_female 0.0094 0.006 1.624 0.108 -0.002 0.021

infant\_mortality\_per1000\_births -0.0112 0.007 -1.578 0.118 -0.025 0.003

per\_capita\_gdp\_dollars 1.5e-05 4.22e-06 3.554 0.001 6.6e-06 2.34e-05

population\_density -9.212e-05 7.08e-05 -1.301 0.197 -0.000 4.87e-05

safe\_drinking\_water\_access\_pct 0.0098 0.005 1.979 0.051 -5.31e-05 0.020

seats\_held\_by\_women\_pct 0.0125 0.006 2.003 0.048 8.64e-05 0.025

unemployment\_rate -0.0373 0.015 -2.555 0.012 -0.066 -0.008

==============================================================================

Omnibus: 6.818 Durbin-Watson: 1.853

Prob(Omnibus): 0.033 Jarque-Bera (JB): 6.580

Skew: -0.494 Prob(JB): 0.0373

Kurtosis: 3.871 Cond. No. 4.26e+05

==============================================================================

Notes:

[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

[2] The condition number is large, 4.26e+05. This might indicate that there are

strong multicollinearity or other numerical problems.