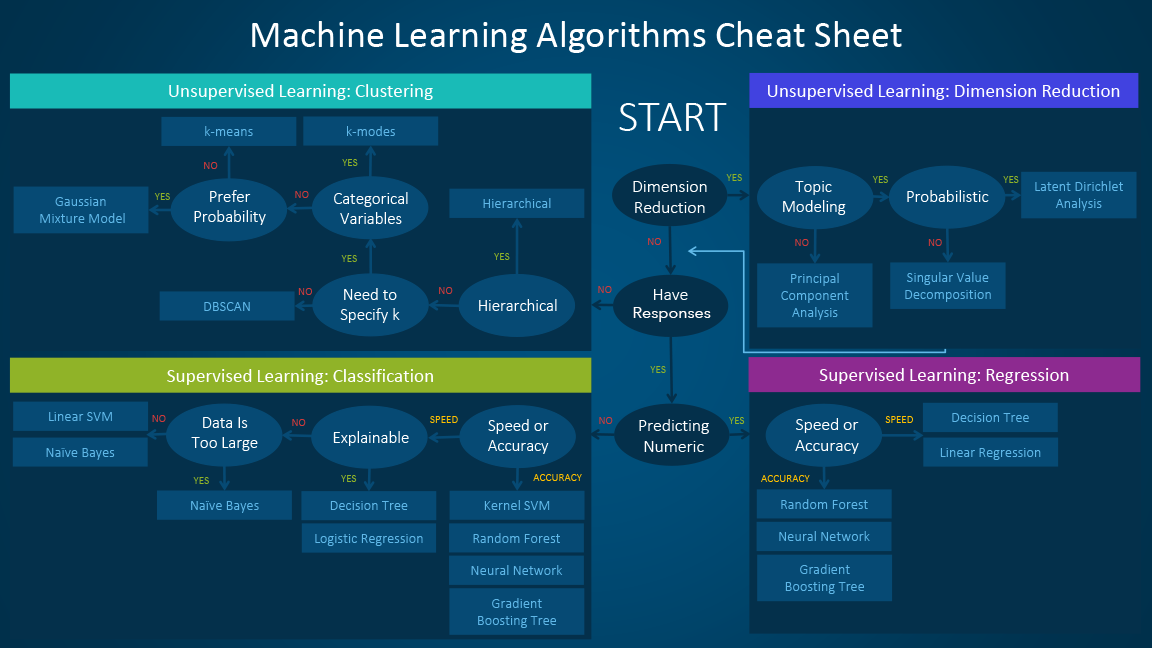
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| S.No | MODEL | WHAT IT DOES | EXAMPLE | CL/REG |
| 1 | Logistic Regression | * Used For Prediction and Classification Problems * A Statistical Analysis Method to Predict a Binary Outcome, Such as Yes or No | * Whether a political candidate will win or lose an election. * A high school student will be admitted or not to a particular college. | Classification |
| 2 | Naive Bayes | * Assumes that the presence of a particular feature in a class is unrelated to the presence of any other feature. * Suitable for binary and multiclass classification and categorical data. | * A fruit may be an apple if it is red, round, and 3 inches wide. * Sentiment Analysis * Spam Filtering | Classification |
| 3 | K-Nearest Neighbour Algorithm | * Stores the dataset and when it gets new data, it classifies that into a category that is like the new data. | * We have an image of a creature that looks like cat and dog, but we want to know either it is a cat or dog. | Classification |
| 4 | SVM (Support Vector Machine) | * Uses classification algorithms for two-group classification problems. | * Handwriting recognition, * Intrusion and face detection * Email and gene classification | Classification |
| 5 | Decision Tree | * Used for classification and regression. * Mimic human-level thinking. * Data is continuously split according to a certain parameter. The tree can be explained by two entities, namely decision nodes and leaves. | * Choose between manufacturing item A or item B * Investing in choice 1, choice 2, or choice 3. | Classification |
| 6 | Random forest | * Classification algorithm consisting of many decisions trees. | * Images of dogs and cats, classifier would predict whether image is of a dog or a cat. * Classifying an email is “spam” or “not spam” | Classification |



**USE make\_classification WITH TRAIN AND TEST SETS FOR BETTER ACCURACY AND CHANGE THE PARAMETERS ACCORDINGLY.**

**EXAMPLE:**

**x\_train1, y\_train1 = make\_classification(n\_samples=1000, n\_features=20, n\_informative=15, random\_state=2)**

**x\_test1, y\_test1 = make\_classification(n\_samples=1000, n\_features=20, n\_informative=15, random\_state=2)**

[**https://scikit-learn.org/stable/modules/generated/sklearn.datasets.make\_classification.html**](https://scikit-learn.org/stable/modules/generated/sklearn.datasets.make_classification.html)

**REGRESSION**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| S.No | MODEL | WHAT IT DOES | EXAMPLE | CL/REG |
| 1 | Linear Regression | * Used to predict the value of a variable based on the value of another variable. * The variable you want to predict is the dependent variable. * The variable you are using to predict the other variable's value is the independent variable. | * Used in business to evaluate trends and make estimates or forecasts. * Sales | Regression |
| 2 | Random Forest Regressor | * A supervised learning algorithm that uses ensemble learning method for regression. * Estimating missing data and maintains accuracy when a large proportion of the data are missing. | * Predict future revenue. * Predict future prices/costs. * Compare performance. | Regression |
| 3 | Multiple Linear Regression | * Used to explain the relationship between one continuous dependent variable and two or more independent variables. | * Scientists use diff. amounts of fertilizer & water on diff. fields and see how it affects crop yield. | Regression |
| 4 | Bayesian Regression | * Has a mechanism to survive insufficient data/ poorly distributed data. * It allows you to put a prior on the coeff. & on the noise so that in the absence of data, the priors can take over. | * Credit card fraud detection. * Spam filtering. | Regression |
| 5 | Decision Tree | * Used for classification and regression. * Mimic human-level thinking. * Data is continuously split according to a certain parameter. The tree can be explained by two entities, namely decision nodes and leaves. | * Choose between manufacturing item A or item B * Investing in choice 1, choice 2, or choice 3. | Regression |

**USE make\_regression WITH TRAIN AND TEST SETS FOR BETTER ACCURACY AND CHANGE THE PARAMETERS ACCORDINGLY**

**x\_train1, y\_train1 = make\_regression(n\_samples=6435, n\_features=9, n\_informative=20, noise=0.5, random\_state=2)**

**x\_test1, y\_test1 = make\_regression(n\_samples=6435, n\_features=9, n\_informative=20, noise=0.5, random\_state=2)**

[**https://scikit-learn.org/stable/modules/generated/sklearn.datasets.make\_regression.html**](https://scikit-learn.org/stable/modules/generated/sklearn.datasets.make_regression.html)

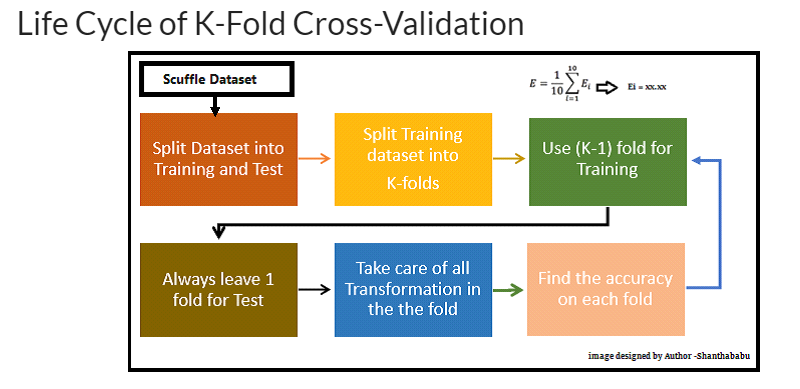
[**https://www.codespeedy.com/make\_regression-function-in-sklearn-with-python/**](https://www.codespeedy.com/make_regression-function-in-sklearn-with-python/)

**Pre-processing**

**Why Kfold for splitting the dataset?**

Ans: K-Folds cross-validator

* Provides train/test indices to split data in train/test sets. Split dataset into k consecutive folds (without shuffling by default).
* It helps us to avoid overfitting (**major advantage**).



* Can be used to divide that datasets so that we will not have imbalanced dataset (use startifiedkfold(classification only) for splitting or kfold)
* Can be used to validate the models (use **cross\_val\_score**) or for Model selection(important).