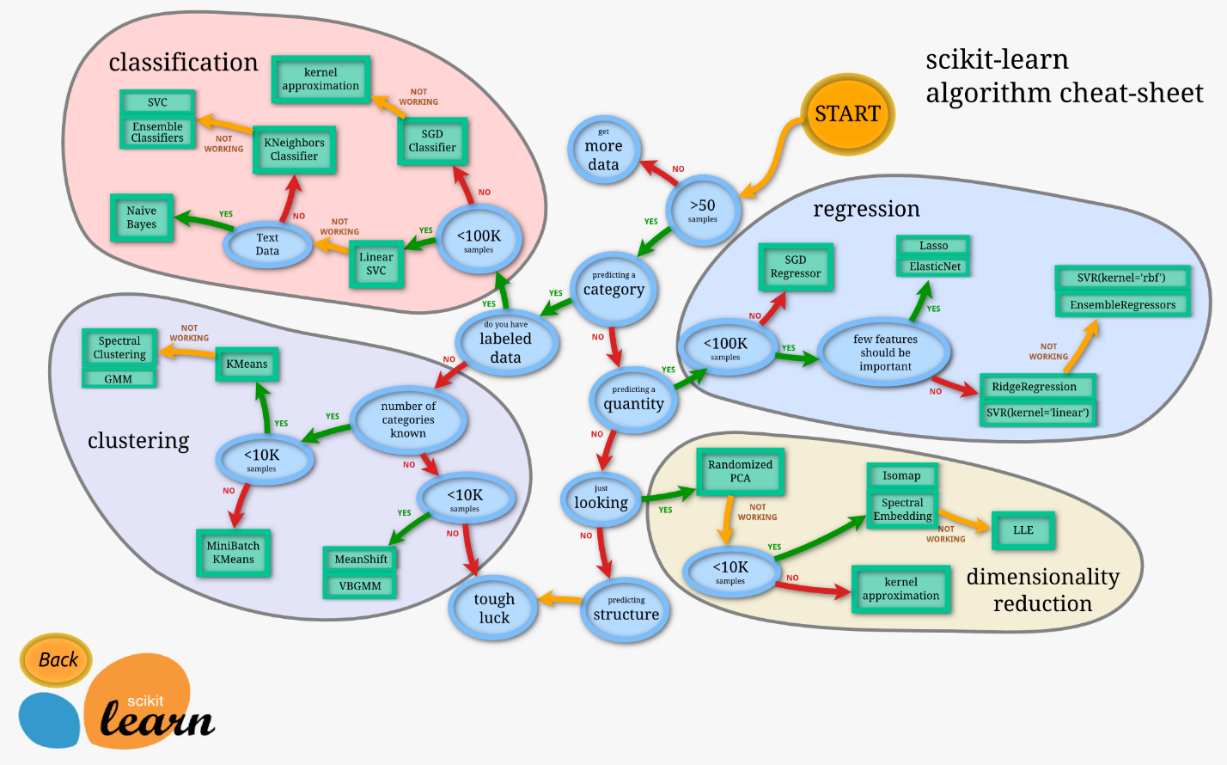
Machine Learning Algorithms

## **The Algorithm Cheat Sheet**



To compare the goodness of fit for different models, we can analyze different metrics like statistical significance of parameters, R2, Adjusted r2, AIC, BIC, error term, and Mallow’s Cp criterion (comparing the model with all possible submodels (or a careful selection of them).

# **Supervised Learning**

## **Generalized Linear Models**

Linear model specifies the (linear) relationship between a dependent (or response) variable Y, and a set of predictor variables, the X's, so that

### **The Many Standard GLM Models**

#### OLS

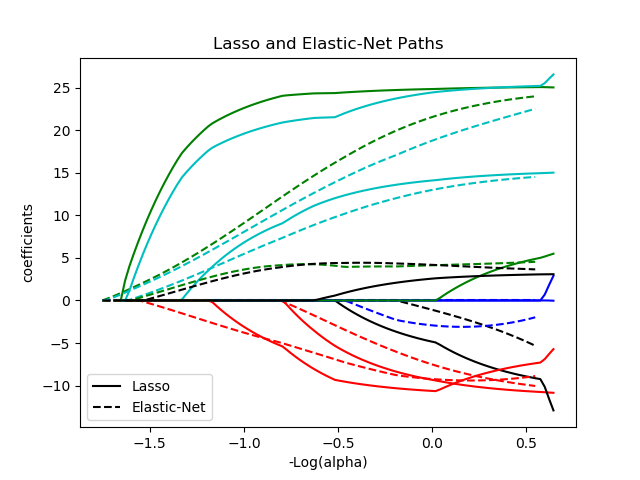
#### Ridge Regression - Imposing a penalty on the size of coefficients which minimize a penalized residual sum of squares. Larger the value of penalizer, the greater the amount of shrinkage and the coefficients become more robust to collinearity of features. L2 regularized (penalty impacts by square of coefficient; use cross val to optimize)

#### Variant HuberRegressor is different to Ridge because it applies a linear loss to samples that are classified as outliers

#### Lasso - Estimates sparse coefficients, it is useful due to its tendency to prefer solutions with fewer parameter values. L1 regularized

#### ElasticNet - Trained with L1 and L2 prior as regularizer. Combination allows for learning a sparse model where few of the weights are non-zero like Lasso, while maintaining the regularization properties of Ridge (trade-off between the two)

Figure: Coefficient paths for various alphas



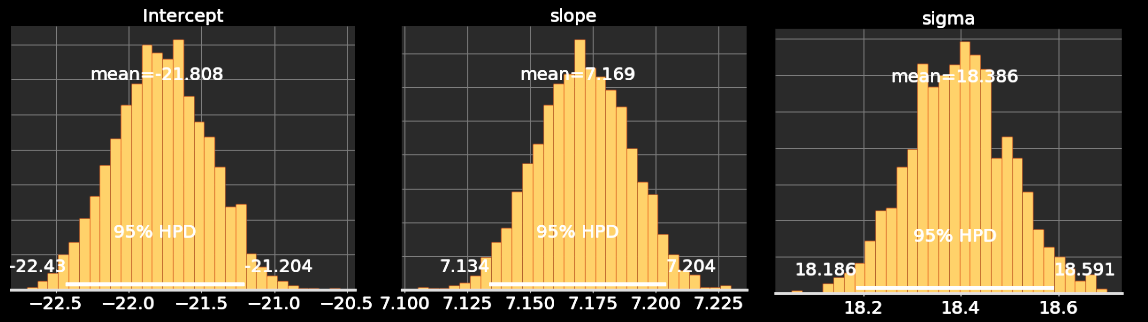
#### Least-angle regression (LARS) – For high-dimensional data, **similar to forward stepwise regression**, at each step, it finds the predictor most correlated with the response (if two variables are almost equally correlated with the response, then their coefficients should increase at approximately the same rate). It is computationally comparable to OLS

#### Stepwise Regression - Successively adding or removing variables based on t-statistics of coefficients. Used for sifting through large dimensions and/or fine-tuning. Can begin with 0 variables and proceed forward or other way round. Inherently used for high dimensional data

#### Orthogonal Matching Pursuit (OMP) - Forward feature selection method like LARS. OMP is based on a greedy algorithm that includes at each step the atom most highly correlated with the current residual

#### Bayesian Regression - Formulate linear regression using probability distributions rather than points. Y is not a single value, but is assumed to be drawn from a probability distribution. Advantage is adapts to the data and include regularization parameters in the estimation

Figure: Coefficients are also derived from Gaussian functions



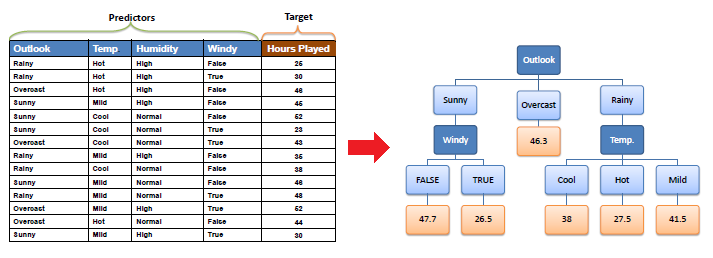
*\*Note that variability increases (larger std) with fewer data!*

#### Polynomial regression - Extending linear models with basis functions; linear models trained on nonlinear functions of the data. Use with caution due to its tendency to overfit

## **Trees and Forest**

### **Decision Trees**

It breaks down a dataset into smaller and smaller subsets. A decision node (e.g., Outlook) has two or more branches (e.g., Sunny, Overcast and Rainy), each representing values for the attribute tested. Leaf node (e.g., Hours Played) represents a decision on the numerical target. The topmost decision node in a tree which corresponds to the best predictor called root node. Decision trees can handle both categorical and numerical data.



#### Algorithms:

1. ID3 (Iterative Dichotomiser) - uses a top-down, greedy search through the given sets, where each attribute at every tree node is tested to select the attribute that is best for classification. Entropy based
2. C4.5 - variant of ID3, better for continuous vars. Uses Entropy based Gain Ratio
3. CART - binary splitting of the attribute, Gini Index is used as splitting attribute
4. Other - CHAID, MARS (numerical data), Conditional Inference Trees

Attribute selection can be done using: Entropy (Information Gain), Gain Ratio and Gini Index (or impurity).

#### Pros and Cons:

|  |  |
| --- | --- |
| **Advantages** | **Disadvantages** |
| Quite interpretable (decision making) | Tend to overfit |
| Multi output | Sensitive to data variance – remedy with bagging/boosting |
| Out-of-the-box | Stuck in local optima |
|  | Biased trees for dominant classes |

<https://www.saedsayad.com/decision_tree_reg.htm>

<https://medium.com/deep-math-machine-learning-ai/chapter-4-decision-trees-algorithms-b93975f7a1f1>

Advantages/disadvantages of decision trees: <https://pdfs.semanticscholar.org/fd39/e1fa85e5b3fd2b0d000230f6f8bc9dc694ae.pdf>

# Introduction

Blockchain technology is disrupting the way we store and access data. Its unique attributes of traceability and immutability are proving to be a disruptive force across a variety of sectors – from retail to manufacturing to healthcare. With much literature exploring the benefits of blockchain implementation in a global context, we assess the benefits of this technology in Sri Lanka by examining case studies, strategies, and test projects in similar emerging and frontier markets. Subsequently, we analyze three key sub-sectors – agriculture, banking and finance, and manufacturing – to look at how blockchains can speed up processes, reduce transaction time and improve logistics performance, and drive cost savings. While doing so, we also examine the areas in which Sri Lanka lags behind the competition and what it should do to bridge this gap.

# Chapter 1: An Introduction to Blockchain Technology

Blockchain is a new type of database system that enables multiple incumbents to store and access the same data in a decentralized manner. Unlike most centralized systems, where data is stored at a central location, blockchain is a network of databases that enables participants to create, disseminate and store information efficiently and securely. Moreover, a blockchain records every transaction or exchange, generating a large base of timestamped information. As such, blockchain technology enables efficient tracking, tracing, and auditing.

## Blockchain: A New Way of Envisioning Databases

A blockchain’s most prominent advantage is its immutability. As blockchains are essentially decentralized databases, where every record is co-located with every participant, also known as nodes, immutability is built into the system. Essentially, altering data in a record would require every record held at each node in the network to be altered. If only a ‘copy’ at one of the nodes is tampered with, the consensus of the rest of the nodes would not accept the tampered record. Thus, blockchains are far more difficult to manipulate than current centralized databases.

**Ensemble Trees**

**Random Forests (Bagging ensemble)**

Ensemble of decision trees that considers a random subset of features when forming questions and only has access to a random set of the training data points. This increases diversity as is known as **Bagging** (aka bootstrap sample). The prediction requires an average of all the individual decision tree estimates. Bias slightly increases but variance decreases. Extremely Randomized Trees are a variant.

<https://medium.com/@williamkoehrsen/random-forest-simple-explanation-377895a60d2d>

### Gradient Tree Boosting

GBRT) is a generalization of boosting to arbitrary differentiable loss functions. GBRT is an accurate and effective off-the-shelf procedure that can be used for both regression and classification problems.

|  |  |
| --- | --- |
| **Advantages** | **Disadvantages** |
| Natural handling of data | Scalability as process is sequential |
| Robust to outliers |  |

\*You can also use general methods such as adaptive boosting and voting classification to build custom trees

## **Linear and Quadratic Discriminant Analysis**

<http://uc-r.github.io/discriminant_analysis>