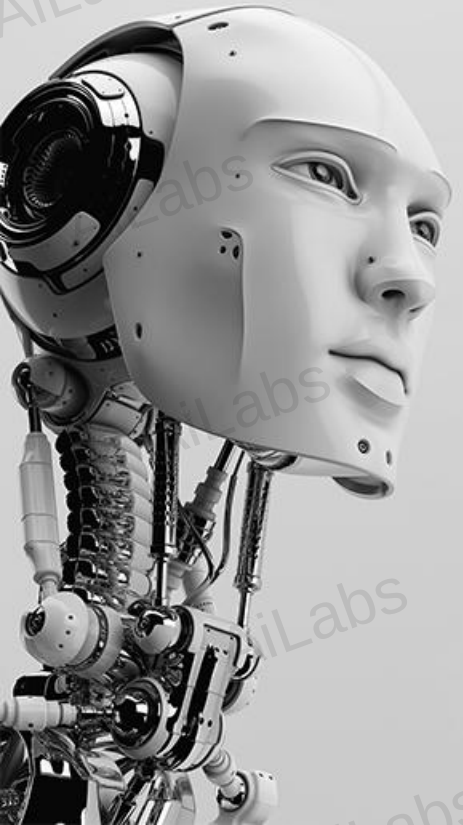
AiLABS

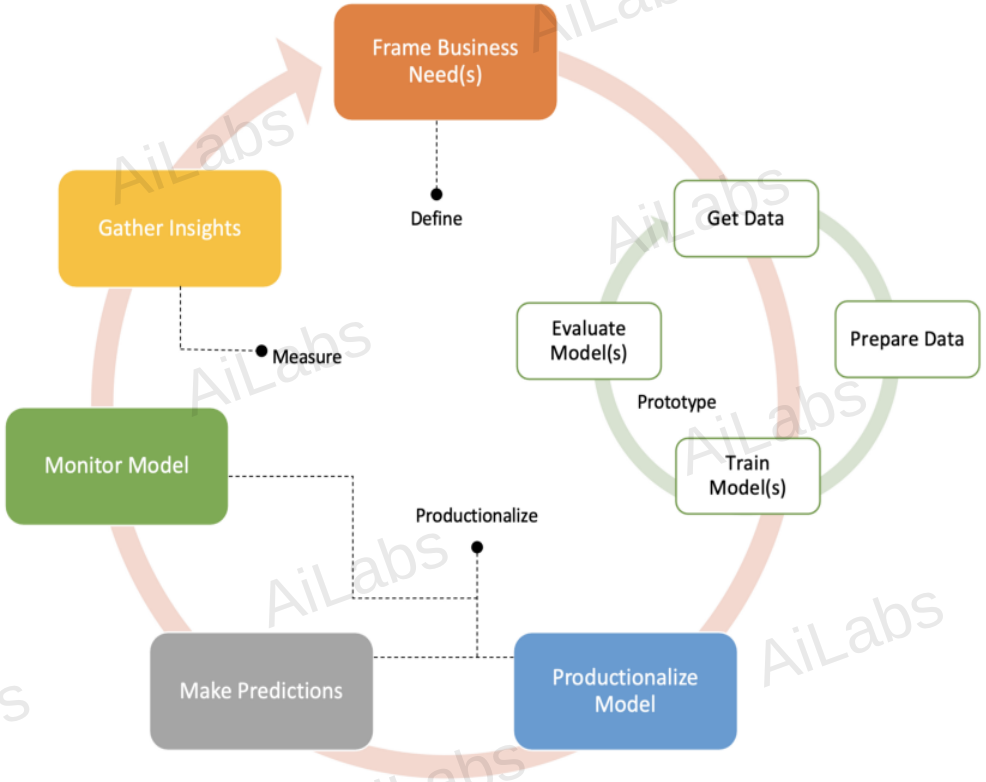


Machine Learning

A Project Lifecycle

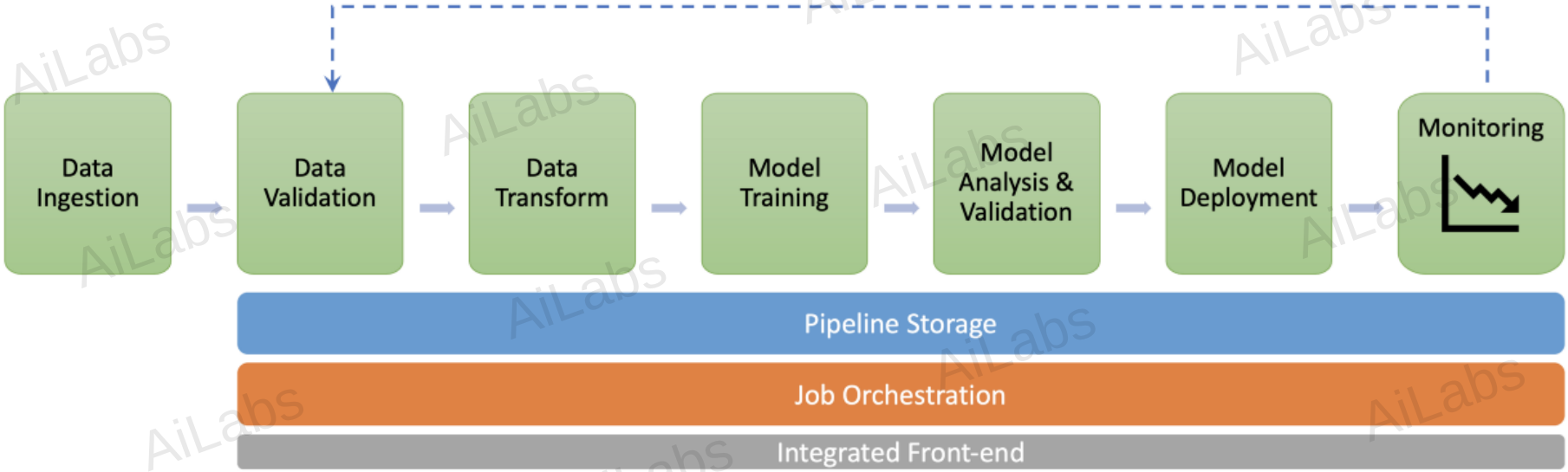
ML Project Lifecycle

A **typical ML lifecycle** can be summarized with the following diagram mainly composed of **3 phases**.



ML Application: Principal Components

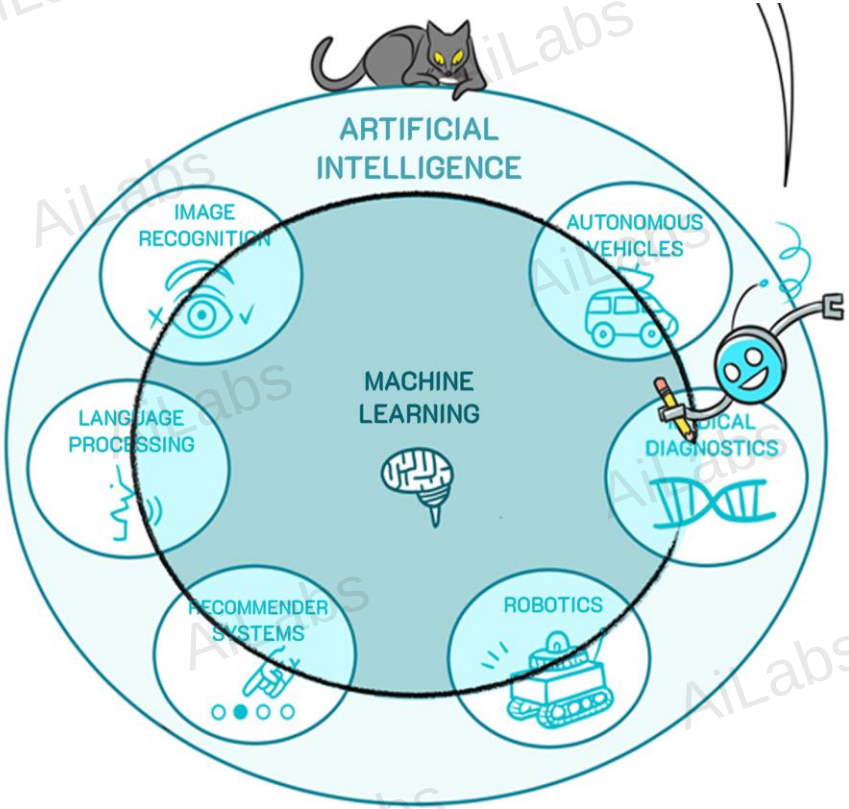
The process of delivering an integrated ML system and continuously operate it in production involves the following steps:



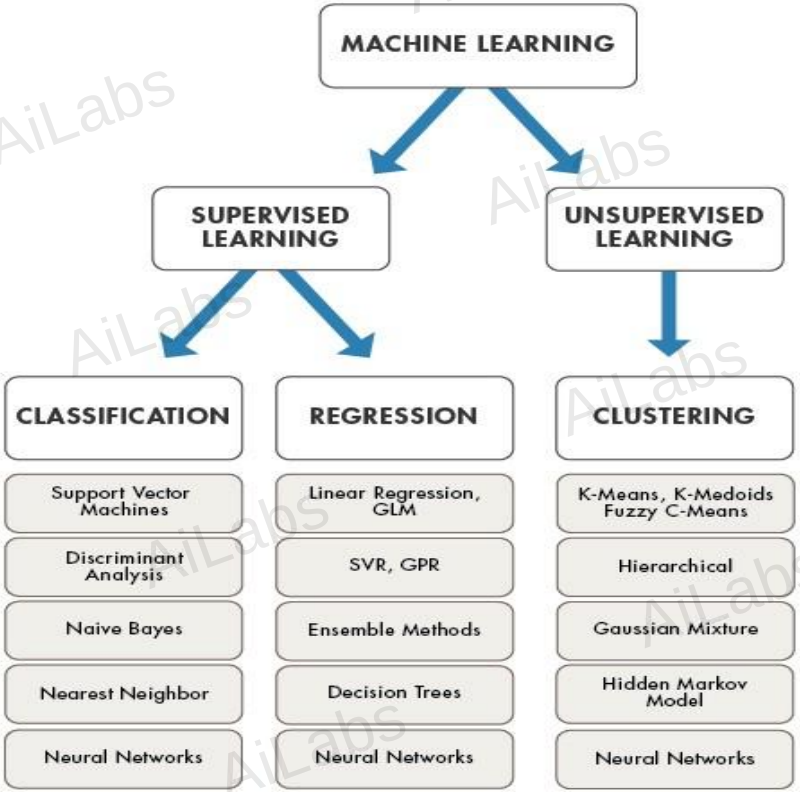
Machine Learning: An introduction

Machine Learning can be broadly classified into Supervised, Unsupervised and Reinforcement Learning.

Supervised learning can be further classified into Classification and Regression.



Machine Learning: Algorithms

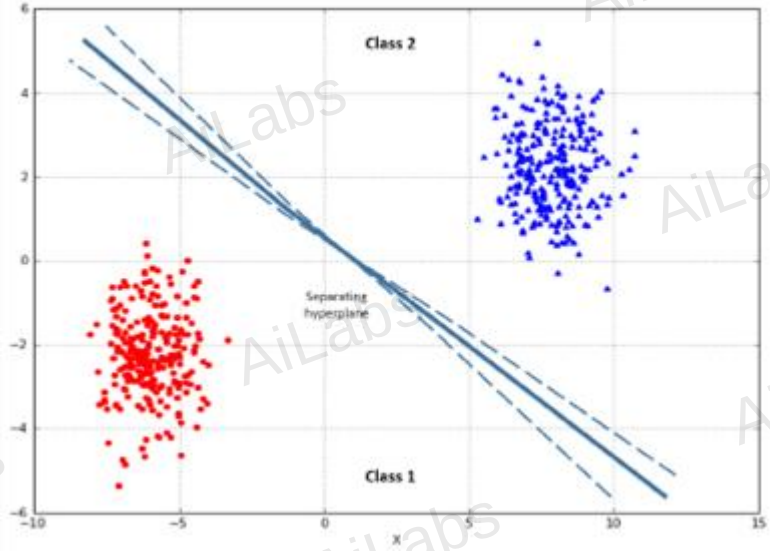


Classification

It is the process of **separating/classifying** two or more types of data **into separate categories** or classes based on their characteristics.

The **output values are discrete** in nature (eg. 0, 1, 2, 3, etc) and are known as “Classes”.

Here, the two classes (red and blue colored points) are clearly separated by the line(s) in the middle. This is an example of classification.



Types of Classification Algorithms

- Linear Models
  - Logistic Regression
  - Support Vector Machines
- Nonlinear models
  - K-nearest Neighbors (KNN)
  - Kernel Support Vector Machines (SVM)
  - Naïve Bayes
  - Decision Tree Classification
  - Random Forest Classification
  - Neural Network

Logistic regression

Logistic regression is a classification algorithm used to assign observations to a discrete set of classes. It is used when dependent variable is categorical.

For example,

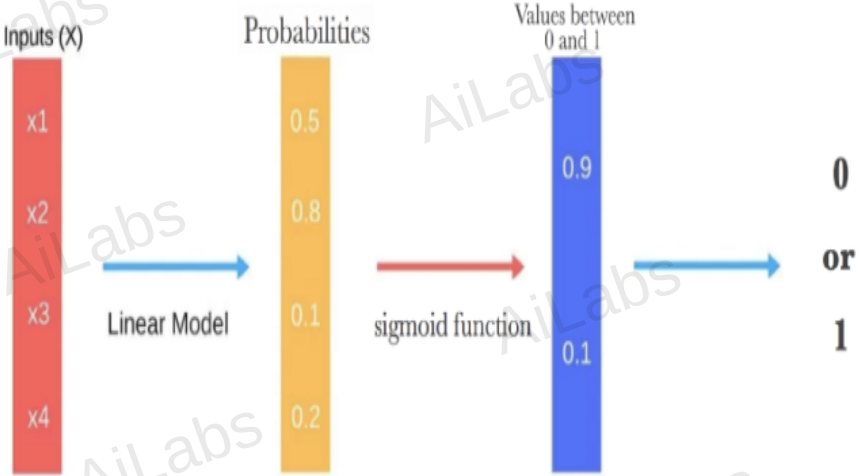
- **To predict whether an email is spam(1) or not(0)**
- **Whether a tumor is malignant(1) ot not(0).**

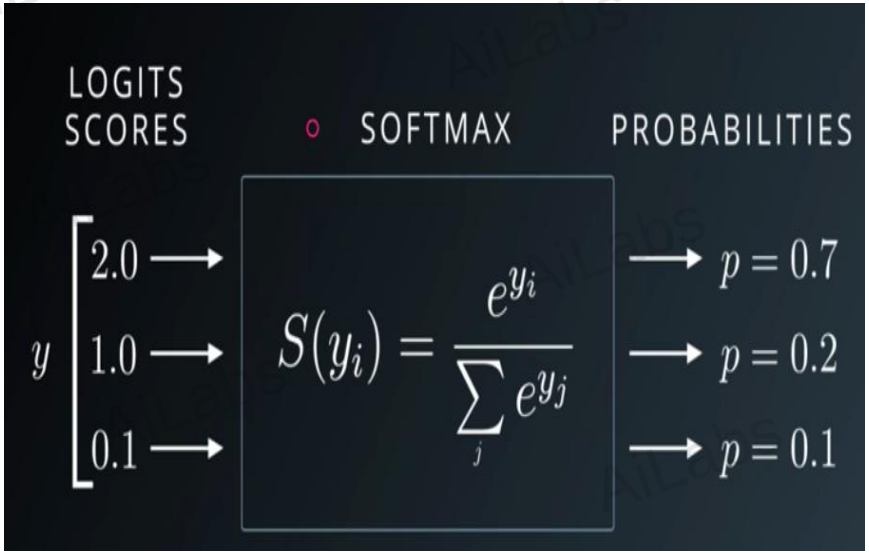
Logistic regression predicts categorical outcomes (binomial / multinomial values of y)

The predictions of Logistic Regression are in the form of probabilities of an event occurring, ie the probability of y=1, given certain values of input variables x. Thus, the results of Logistic Regression range between 0-1.

Logistic Regression

Sigmoid and Softmax activation functions in Logistic regression.

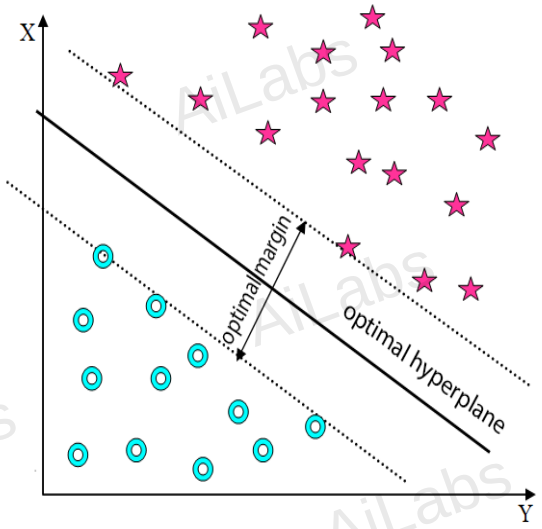




Support Vector Machine

- Support vector machine draws a hyperplane in n dimensional space such that it maximizes the margin between the classifying groups.

- The dimension of the hyperplane depends upon the number of features. If the number of input features is 2, then the hyperplane is just a line. If the number of input features is 3, then the hyperplane becomes a two-dimensional plane. Support vectors are data points that are closer to the hyperplane and influence the position and orientation of the hyperplane. Using these support vectors, we maximize the margin of the classifier.



K-Nearest neighbour

- K nearest neighbors is a simple algorithm that stores all available cases and classifies new cases based on a similarity measure (e.g., distance functions - Euclidean, Manhattan etc)

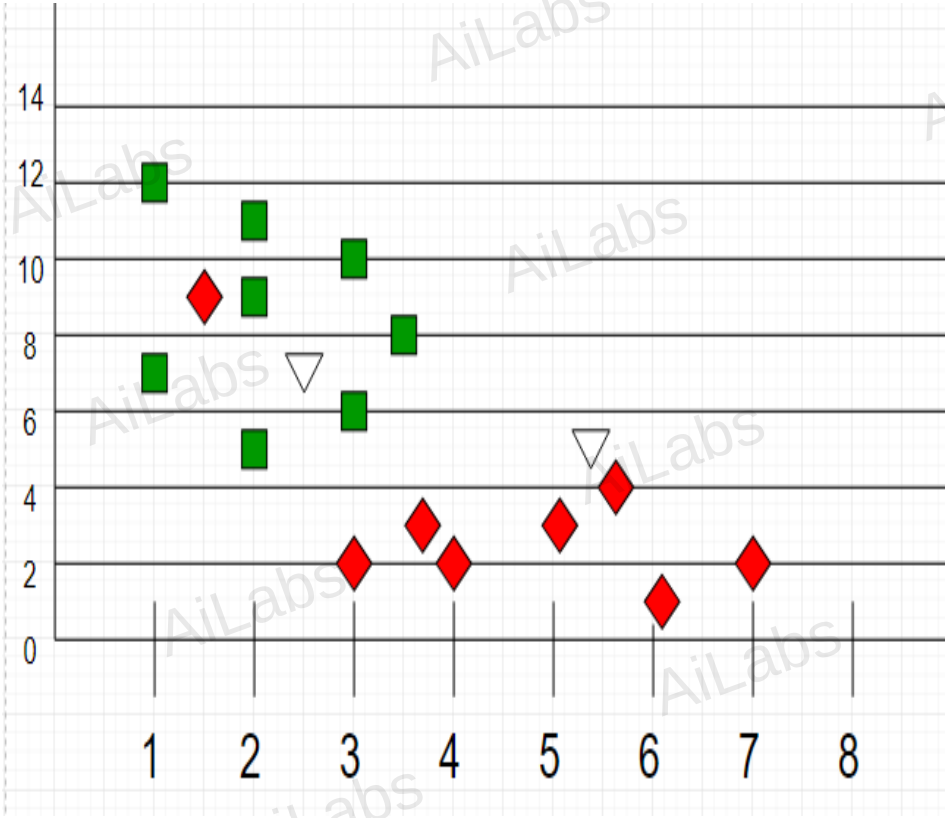
- The KNN algorithm assumes that similar things exist in close proximity. In other words, similar things are near to each other.

  “Birds of a feather flock together.”

- A case is classified by a majority vote of its neighbors, with the case being assigned to the class most common amongst its K nearest neighbors measured by a distance function. If K = 1, then the case is simply assigned to the class of its nearest neighbor.

K-Nearest neighbour

Given an unclassified point, we can assign it to a group by observing what group its nearest neighbors belong to. This means a point close to a cluster of points classified as ‘Red’ has a higher probability of getting class. Intuitively, we can see that the first point (2.5, 7) should be classified as ‘Green’ and the second point (5.5, 4.5) should be classified as ‘Red’.



Naive Bayes

- A Naive Bayes classifier is a probabilistic machine learning model that's used for classification task. The crux of the classifier is based on the Bayes theorem.

- Bayes Theorem

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

- The assumption made here is that the predictors/features are independent. That is presence of one particular feature does not affect the other. Hence it is called naive.

Naive Bayes

Naive Bayes classifier calculates the probabilities for every factor Then it selects the outcome with highest probability.

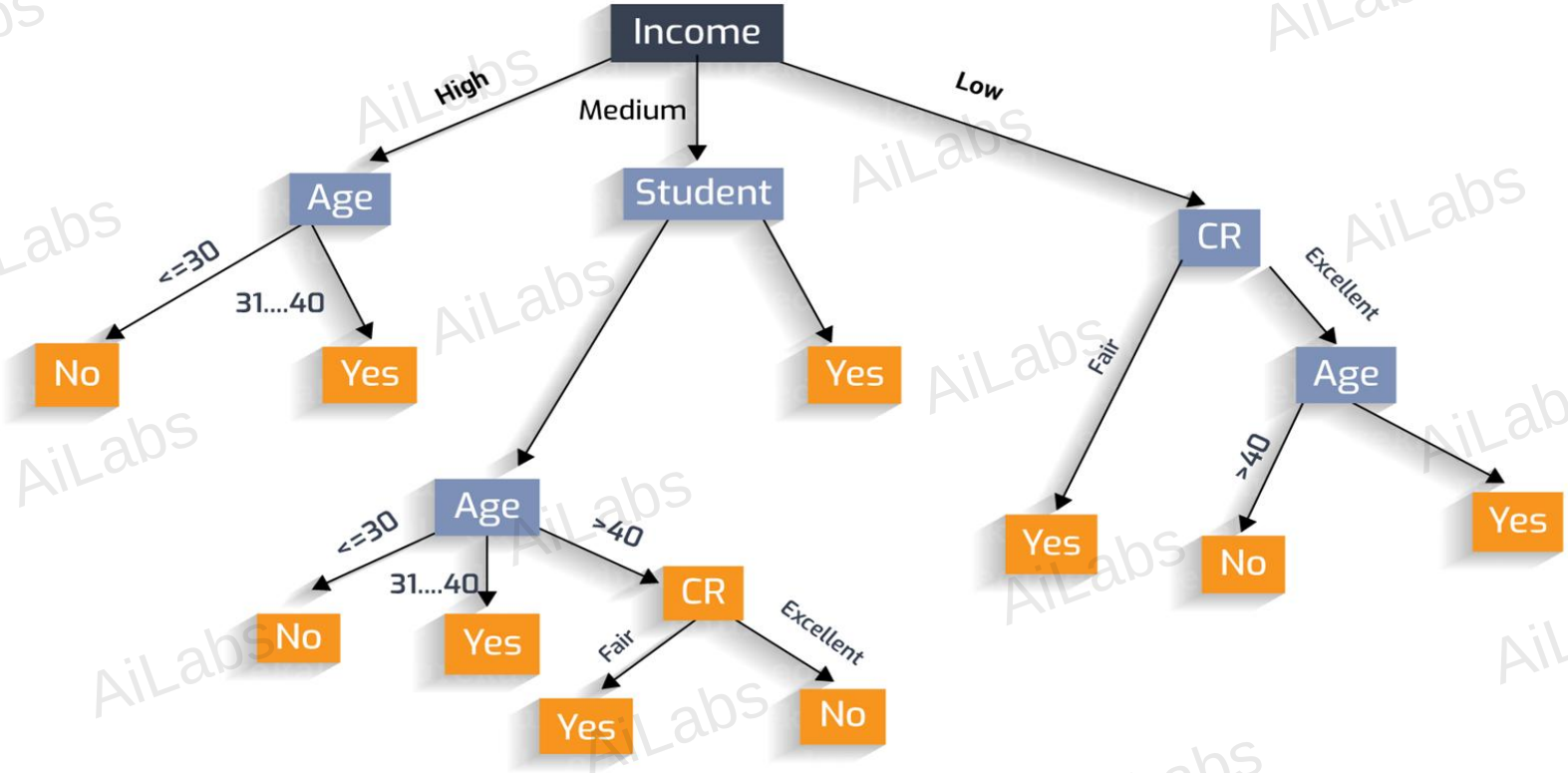
Even with this it is powerful algorithm used for

- ➢ Real time Prediction
- ➢ Text classification/ Spam Filtering
- ➢ Recommendation System

Decision Tree

- A decision tree is a flowchart-like tree structure where an internal node represents feature(or attribute), the branch represents a decision rule, and each leaf node represents the outcome.
- The topmost node in a decision tree is known as the root node and the last nodes are called the leaf node.
- It learns to partition on the basis of the attribute value. It partitions the tree in recursively manner call recursive partitioning.

Decision Tree



Random Forest

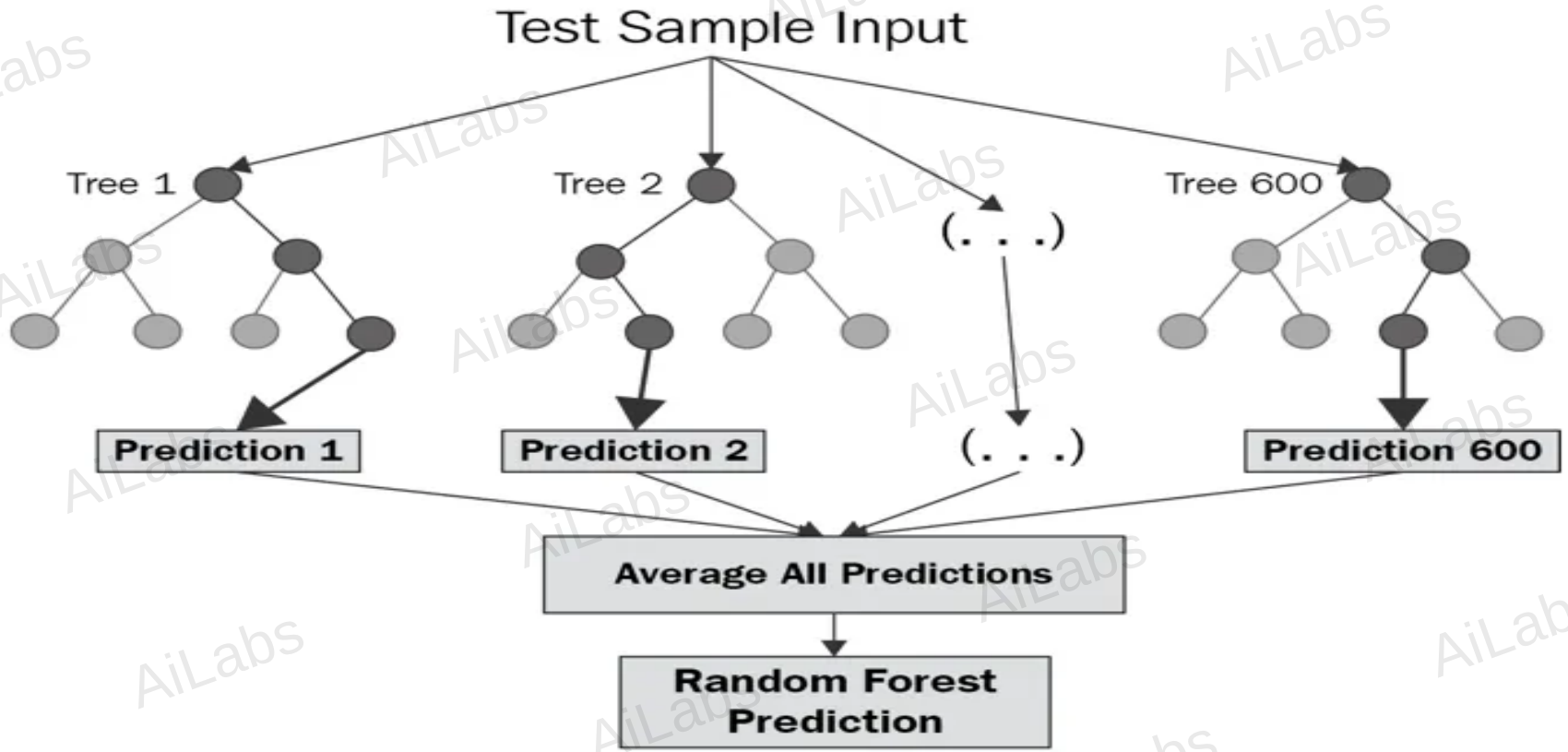
- Random forest is a technique used in modeling predictions and behavior analysis and is built on decision trees. It contains many decision trees that represent a distinct instance of the classification of data input into the random forest.

- The random forest technique takes consideration of the instances individually, taking the one with the majority of votes as the selected prediction.

- Each tree in the classifications takes input from samples in the initial dataset. Features are then randomly selected, which are used in growing the tree at each node.

- Every tree in the forest should not be pruned until the end of the exercise, when the prediction is reached decisively. In such a way, the random forest enables any classifiers with weak correlations to create a strong classifier.

Random Forest

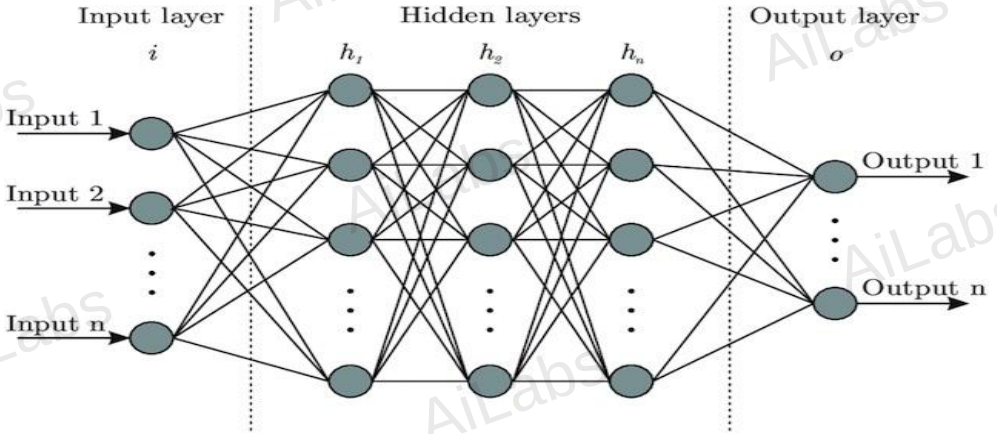


Neural Network

NETWORK LAYER:

- The commonest type of artificial neural network consists of three groups, or layers of units:
- a layer of "**input**" units is connected to a layer of "**hidden**" units, which is connected to a layer of **"output"** units.

Each neuron is connected to each other through a weighted connection.

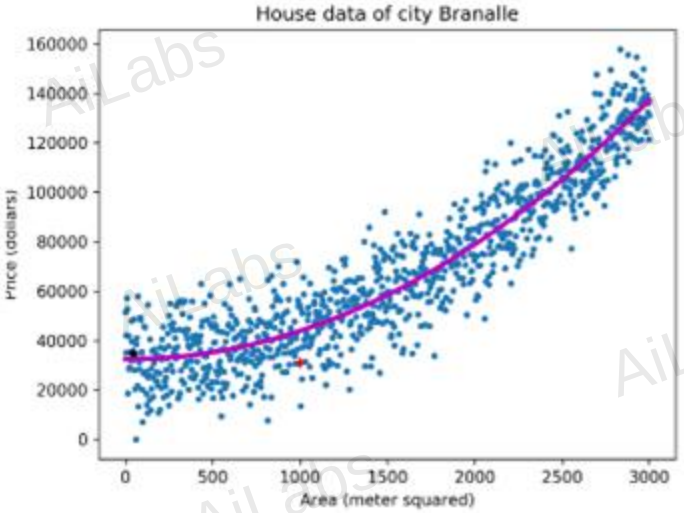


Regression

It is the process of **determining the "Best Fit" curve** for the given data such that, on unseen data, the data points lying on the curve accurately represent the desired result.

The **output values are continuous** in nature (eg. 0.1, 1.78, 9.54, etc).

Here, the curve represented by the magenta line is the "Best Fit" line for all the data points as shown. This is an example of Regression.



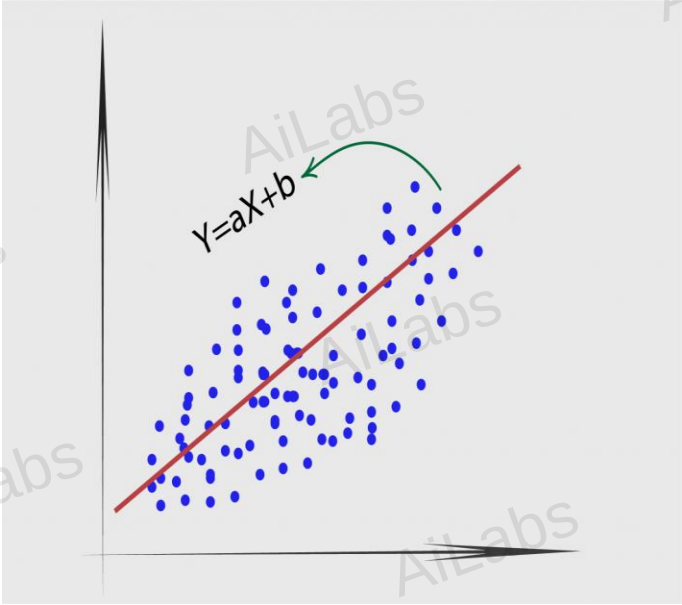
Linear Regression

It establishes a relationship between a dependent and an independent variable using a best fit straight line also known as regression line.

The best fit is achieved using what we call Least Squared Error Method.

It finds the difference between the actual data and the predicted data and find the sum of their squares.
It is given by:

$$Y = a*x + b$$



Polynomial Regression

- In this type of regression the independent and dependent variable doesn't follows a linear relationship. The relationship they follows can only be modeled when the independent variable are raised to a power.

- It follows:

$$Y = a*x^n + b \quad \text{where } n = 2,3\ldots.$$

- One of the widespread problem with this is Overfitting. If we use a very small value of n the model may underfit while if we use a large value of n the model may overfit. Using the right value is the key.

Ridge Regression

- This type of technique is used when data suffers from multicollinearity(Independent variable are highly correlated).Ridge regression reduces the standard error by adding a bias to the variance.

- The equation for ridge regression are:

$$Y = a*x + b+ e$$

- Here ‘e’ is the error term that accounts for error n real and prediction value.Ridge regression solve the multicollinearity problem through shrinkage parameter.

Lasso Regression

Similar to ridge regression lasso regression also penalises the absolute size of the regression coefficient.

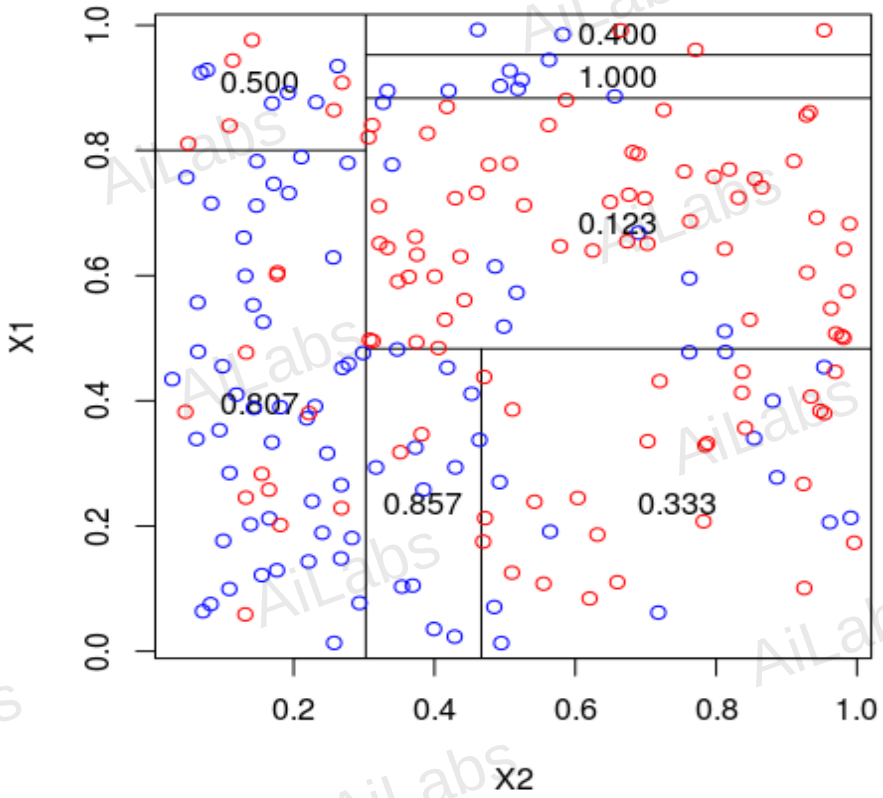
Lasso regression is capable of improving the accuracy of the linear regression model.

From the equation below we see that Lasso regression uses absolute value to penalise the loss function.

This is basically L1 regularisation.

$$= \underset{\beta \in \mathbb{R}^p}{\operatorname{argmin}} \underbrace{\|y - X\beta\|_2^2}_{\text{Loss}} + \lambda \underbrace{\|\beta\|_1}_{\text{Penalty}}$$

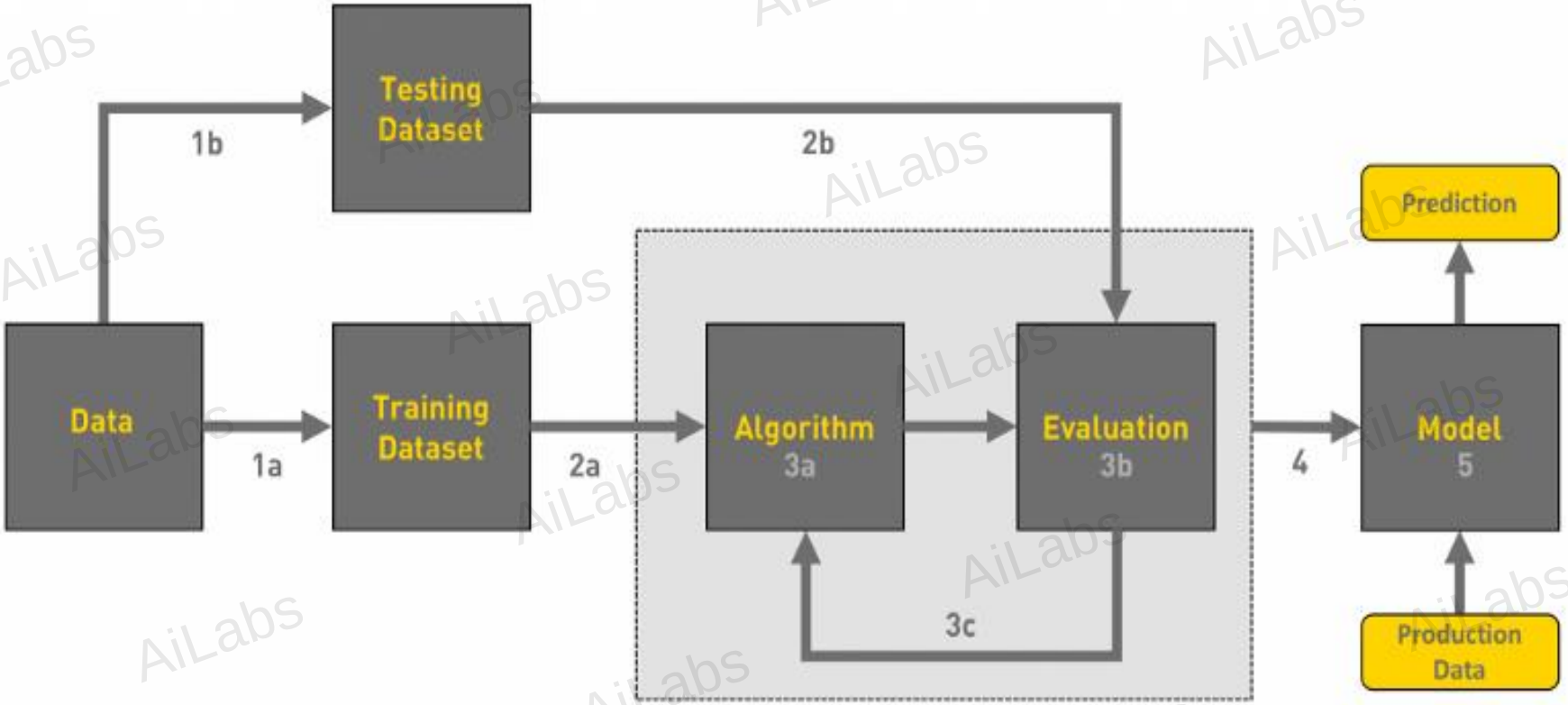
Decision Tree Regression



Workflow of a Machine Learning Project

1. **Gathering data** - The process of gathering data depends on the type of project we desire to make.
2. **Data preprocessing -** It is a process of cleaning the raw data into clean data, so that can be used to train the model. It is needed to achieve good results from the applied model in machine learning projects.
3. **Researching the model that will be best for the type of data -** The goal is to train the best performing model possible, using the pre-processed data.
4. **Training and testing the model on data -** For training a model the data is split into 3 three sections which are ‘Training data’ ,‘Validation data’ and ‘Testing data’.The classifier uses ‘training data set’, tunes the parameters using ‘validation set’ and then test the performance on unseen ‘test data set’. An important point to note is that during training the classifier only has training and/or validation set is available. The test data set must not be used during training the classifier. The test set will only be available during testing the classifier.

Workflow of a Machine Learning Project



Data Preprocessing

Data **preprocessing** is the **most important phase** to improve the quality of the data in order to **improve performance** of the ML application.

The **raw data** obtained from the real-world is likely to contain a **good amount of noise** in it.

The data might **not be homogenous**, which means, the values of different “features” might belong to different ranges.

Hence, after the **removal of noise**, the data needs to be **normalized or scaled** in order to make it homogeneous.

General Steps to follow for Data Preprocessing

Basic workflow of data preprocessing:

1. Drop the unnecessary columns which do not have any visible significance, ex- id,order number,invoice number etc.

1. Look for any kind of outliers and eliminate them.

1. Look for missing values in any column(in our problem Total_Charges had some missing values), after finding the null values, we need see what kind of method can be used to fill the null values.
    a. Delete the rows with missing values - This can be used when the number of missing values is not too high(>10%).
    b. Replace with mean/median/mode - This strategy can be applied on a feature which has numeric data like the age of a person or the ticket fare. We can calculate the mean, median or mode of the feature and replace it with the missing values.

1. Handling the categorical data - the categorical data should be converted to numbers either using label encoding or one-hot encoding

1. Correlation map can be plotted to find out least significant feature, and that feature can be removed from the dataset

1. Feature scaling for data to come down to same range

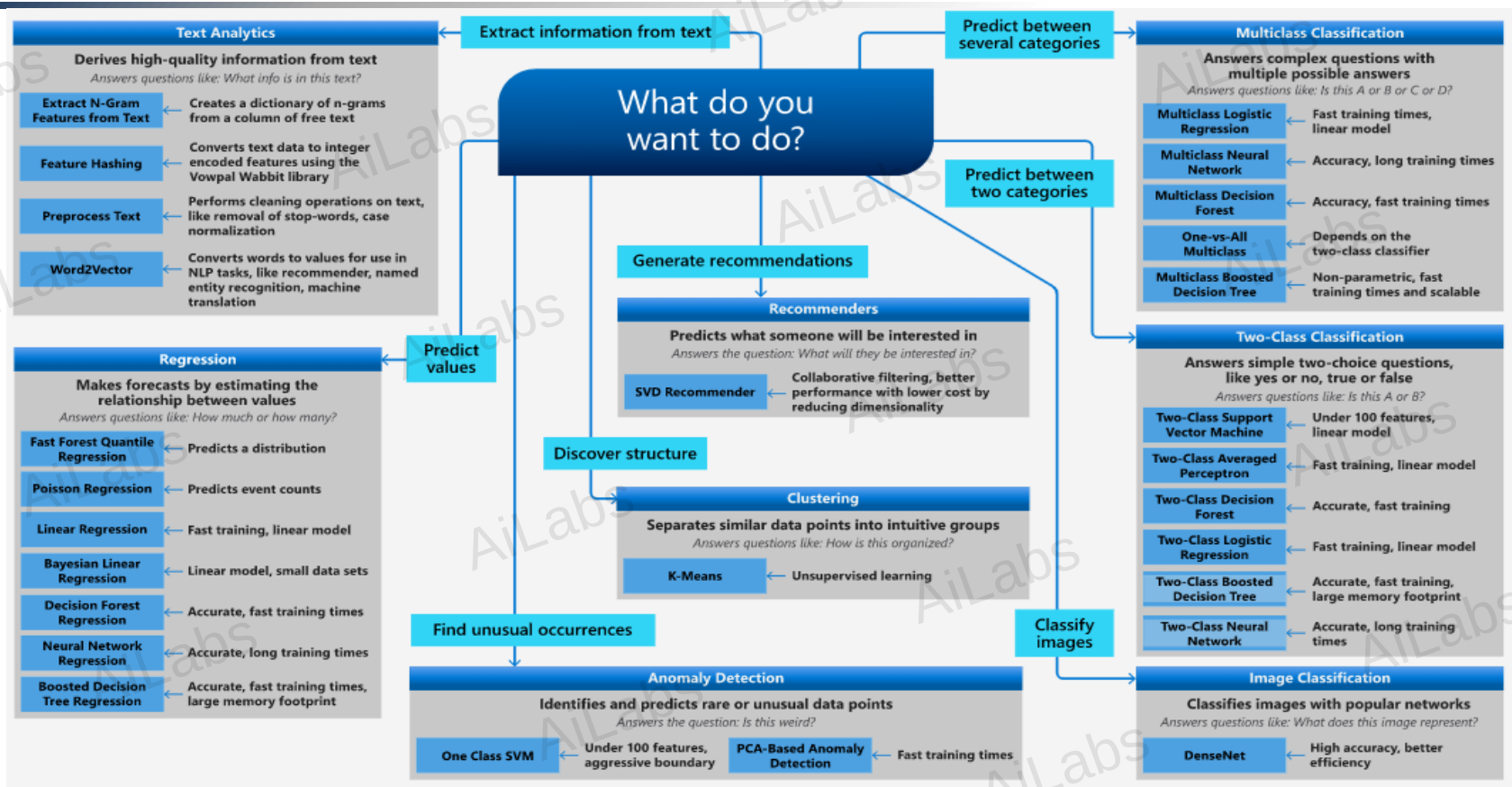
Feature Engineering

In a tabular dataset, all the **columns** that describe the data are called “**Features**”.

Data often contains **features** which are **irrelevant** to the **output of the model**. Hence, these **features need to be removed** or **statistically processed** to make sure that they do not interfere with the training of the model on features that are relevant.

In addition to the removal of irrelevant features, it is often required to **create new relevant features from the existing features**. This allows the model to learn better and this process is also called “Feature Extraction”.

Choose the Right Algorithms



Contact Us

DG Block, Sector II, Plot 4. Salt Lake City Kolkata - 700 091

9339850883/9051324266

info@ailabs.academy