# CHAPTER 2

# INTRODUCTION TO MACHINE LEARNING TECHNIQUES

## 2.1 Introduction

There has recently been a huge increase in data with the extended use of the Internet and digitization. In today's scenario, data is generated every day due to online sales, website traffic, daily transaction, social media, etc. A large amount of data is available for analysis. So, The Machine Learning concept is something that comes from this environment. The basic idea of using Machine Learning is that computers can analyse data to find similar patterns, in ways that are too complex for humans to find manually.

As mentioned in Chapter 1, students are the most important component of the education sector, as the progress and status of any University or Institution is based on its students' academic performance. Usasmah et al. (2013) stated that, through learning assessment and extra-curricular activities, student performance can be measured. Many such studies have been done on academic data of graduation students.

The chapter starts with an introduction to Machine Learning (ML). This is followed by a discussion of classification techniques, which are part of supervised learning, which we will be using for prediction. We also discuss Ensemble methods of ML, which use combinations of individual ML algorithms, for better efficiency.

## 2.2 Machine Learning

Machine learning is the subfield of artificial intelligence (AI) that gives the ability to automatically learn and improve performance from experience, without being explicitly programming the system. ML focuses on the progress of computer programs that can access data and achieving relevant tasks by using it. The learning process begins with observations or data; for example, see patterns in the data and make better decisions based on the examples on unseen data. The main aim of machine learning is to enable computers to learn automatically without human interference or help and regulate activities accordingly.

Arthur Samuel defined ML as "Machine Learning in the field of study that gives computers the ability to learn without being explicitly programmed.'' Further lately, Tom Mitchell of Carnegie Mellon University, in 1997, went further into the engineering

concept and said that “A computer program is said to learn from experience E concerning a task T and some performance measure P, if its performance on T, as measured by P, improves with experience E”.

Ensemble methods combine several base models to produce one optimal predictive model. Ensemble learning helps improve ML outcomes by combining multiple models. These methods allow better analytical performance than a single model (Wezel et al., 2007). They are meta-algorithms that combine ML techniques into one analytical model to decrease variance**,** bias or improve predictions**.** Some methods use different types of learners.

There are three key methods, commonly used to construct ensemble classifiers: bagging, boosting, and stacking. Different authors have examined with theoretical and statistical studies, the primary mechanism of bagging, boosting, and stacking, alongside their variants and highlighting their good performances and their ability to reduce classification errors and bias (Webb et al., 2004).

Bagging is an ensemble learning algorithm that is a short form for bootstrap aggregation. Bagging uses subsets to get a fair idea of distribution. In bagging multiple subsets are made from the original dataset. A base model is created on each of these subsets. Each model executes in parallel and independently. The final result of this method by combining the predictions from all models. It is an easy algorithm to implement and gives good performance results.

Boosting is another ensemble method that is quite different from bagging. It is usually used for improving the performance of any classifier and reducing the error of the weak one. It is a successive process, where each succeeding model attempts to correct the errors of the previous one. In boosting a subset is made from the original dataset. Initially, all the data points are given the same weights. A base model is created on this dataset. Then errors are calculated. The observation s which are wrong classified are given higher weights. In every step, the sample distribution was changed to put more weight on misclassification. The final result is a weighted average of all the weak learners. Similarly, just like bagging, boosting is often applied to the same or similar algorithm and uses majority voting for its decision approach.

Stacking**,** also known as stacked generalization. It is an Ensemble method which is the combination of several methods combined in different ways, by introducing the concept

of a Meta learner. It works to construct several different learners that are used to create an intermediate predicate (output values) that becomes the input for the meta classifier for the final prediction.

## 2.3 Classification Techniques

Classification is the supervised method for classifying the target class accurately, based on the dataset. Classification algorithms are discrete, hence useful for predicting outputs. In other words, they are useful when the answer to a question falls under an approximate set of possible outcomes. The s-shaped mathematical function S(x), called the Sigmoid function, is used as the Predictor function in classification problems.

In classification, the Dataset is split into two parts: training dataset and testing dataset. The training dataset is used to find a certain pattern between the independent and dependent variables. The testing dataset compares the result of the model on the new (unseen) data to evaluate the built model and find its accuracy.

Suhaimi et al., (2019) analysed student's graduation time based on prominent factor by using classification methods. One more study (Roy & Garg 2017) used classification methods for identifying weak students and factors that influence the academic performance of students. Hasan et al., (2019) has used classification techniques to find out the current status and predict student's future. Rahman & Islam (2017) used behavioural attributes and student absence in class. This study applied classification methods on these features to evaluate student performance.

Different individual and Ensemble Methods have been used in student performance classification. The most common techniques are Support Vector Machine, K Nearest Neighbour, Decision Tree, Random Forest, Naïve Bayes, Boosting and Bagging. The techniques have been used to compare and find the best classifier for a given dataset.

These techniques discuss in the following sections.

### 2.3.1 Logistics Regression

Logistic Regression is used to perform binary classification to build the classifier models (Lio & Chin, 2007). A certain function between the target variable, i.e., categorical variable and the independent variables, is calculated by estimating the

probabilities using a logistic function (Dominguez-Almendors et al., 2011). The current study wants to classify students into binary class (Technical taking digit 1 and non-technical taking digit 0). Also, this study classifies based on result into 5 categories failure, poor, satisfaction, very good and excellent into the digit 1 to 5.

A student who has potential to take admission in technical programs will have probability (.5< p <1). A total of 20 variables are used expressed as ($x_1, x_2 \ldots, x_{20}$).

The logistic Regression hypothesis in Equation 2.1 that is defined as (Ng, 2011)

$$h_{Q\,(x)} = g(Q\,T_x) = {}^{1}\Big/\left(1 - e^{-(Q^0 T_x\ )}\right) \qquad (2.1)$$

Where $h_Q(x)$ – probability that output is 1 on x,

Q – parameters that used to be fitted to the data

S – sigmoid function, it is shown in Equation 2.2

$$S = \frac{1}{(1 - e^{-2})} \qquad (2.2)$$

This model using all the 20 attributes used in this study is shown in Equation 2.3

$$h_Q\,(x) = S\,(Q_0 + Q_1\,x_1 + Q_2\,x_2 + \cdots Q_{20}\,x_{20}) \qquad (2.3)$$

Where S is the sigmoid function,

Q – parameters selected from the training set

($x_1, x_2 \ldots, x_{20}$) are the 20 attributes used in this study. This equation shows the boundary that divided that student in the technical or non-technical program.

$$I\,(\Theta) = \frac{1}{n}\left[\sum_{i=1}^{n} y^{(i)} + (1 + y(i)\ \log(1 - h_Q\,(x^{(i)})\right] + \frac{y}{2^n}\ \sum_{I=1}^{m} \Theta_I^2 \qquad (2.4)$$

Where $I\,(\Theta)$ is the cost and Θ is the vector of the parameters of the training set.

To determine the parameters that minimize I (Θ) in eq for finding the parameters Q. Through this parameter was tuned to achieve logistic regression model from the given input-output data as shown in Figure 2.1 (Wong & Chen, 1999).

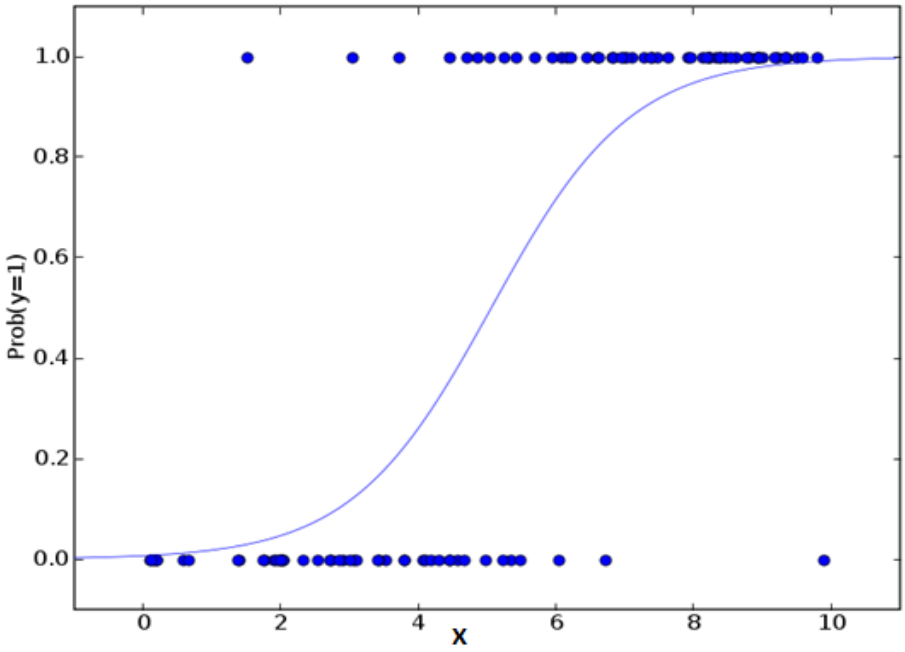


**Figure 2.1** Logistic Regression

Source: Fernandes, 2020

### 2.3.2 Support Vector Machine

Support Vector Machine (Figure 2.2) is part of a supervised ML algorithm that is used for solving classification or regression problems. The data is classified into different classes by searching a line (hyperplane) that separates the training dataset into different classes. This method plots each data value as a point in an n-dimensional space, where n is the number of attributes, with the value of each attribute being the value of a particular coordinate. Those points which are closer to the hyperplane are called support vectors. We try to maximize the margin of the classifier by using support vectors as given in Figure 2.2.

SVM is used to carry out binary classification (Tang, 2013). This model is suitable for use in our study, where training data output is expressed as ($x_i$, $y_i$), where (i = 1,2,…..,m) and $y_i$ € {1,0}; 1 represents a student able to take admission in a technical program and 0 represents the student who took admission in a non-technical program. This theory and the equation of SVM were used from lecture notes (Nag, 2012).

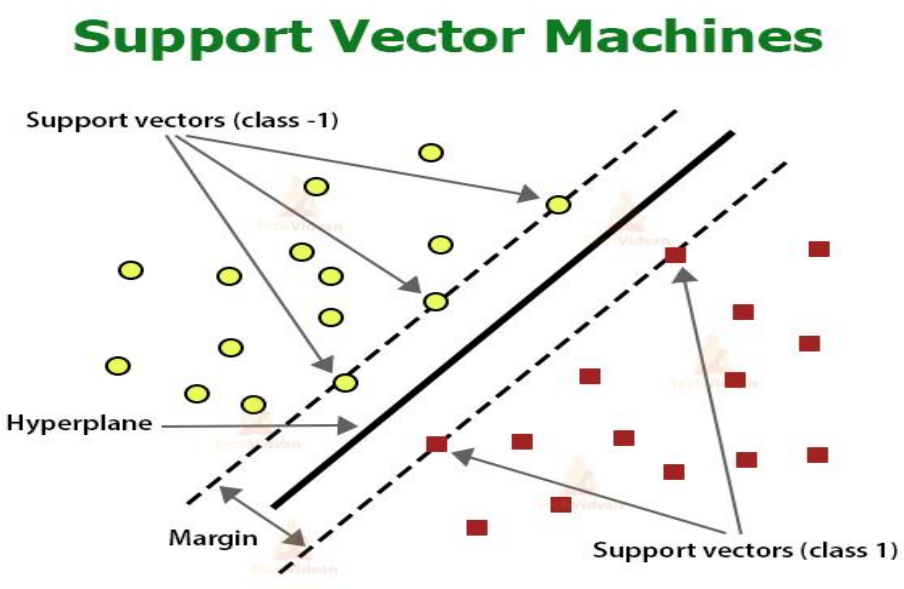


**Figure 2.2** Support vector Machine

Source: Techvidvan, 2021

The Equation 2.5 is derived from the SVM decision boundary:

$$min_{\Theta}^{T} \sum_{i=1}^{m} [y^{i}\ cost, \left(Q^{T}x^{(i)}\right) + \left(1 - y^{(i)}\right) cost_{0}\ (Q^{T}x^{(i)})] + \frac{1}{2} \sum_{i=1}^{n} \Theta_{j}^{2} \qquad (2.5)$$

Where T is the regularization parameter that defines the line that separates between positive and negative samples.

SVM has the capability to separate samples with as large a margin as possible.

### 2.3.3 Naïve Bayes

This method is a classification method which uses Bayes Theorem. It is an assumption of independence between predictors. In other words, an NB classifier predicts that the presence of a particular attribute in a class is not related to the presence of another attribute. NB model is easy to construct. It is mainly useful for very large data sets.

The NB classifier is a category of Bayesian classifier. This model assigns the given sample to the most likely class as explained by the attribute vector (Leung, 2007). This classifier assumes each attribute as independent (Murphy, 2006). The process of NB for classification is as follows:

Let there be 20 attributes ($x_1, x_2 \ldots, x_{20}$) as possible samples. X is called “Evidence”.

H is the hypothesis that assigns the evidence X to one of the target classes, say i. The aim is to determine the posterior probability of hypothesis H, given the sample X,. The probability is calculated based on Bayes' theorem as shown in Equation 2.6.

$$P\left(H \mid X\right) = \frac{P\left(X \mid H\right) P(H)}{P(X)} \qquad (2.6)$$

To determine the class that the probability maximizes the highest probability belong to a class Ci. The highest class Ci is expressed as $P(C_i \mid X)$, the maximum posterior hypothesis.

$$P\left(\mathrm{Ci} \mid \mathrm{X}\right) = \frac{P\left(X \mid \mathrm{Ci}\right) P(\mathrm{Ci})}{P(X)} \qquad (2.7)$$

Where P(X) is the prior probability

In Equation 2.7, for reducing the cost of computing, $P(X \mid C_i)$ and so reduce the analysis for $P(X \mid C_i)\ P(C_i)$, there has a strong assumption that the attributes are dependent. This assumption allows for the mathematical expression in Equation 2.8:

$$P(X \mid C_i) = \prod_{k=1}^{n} P(X_k \mid C_i) \qquad (2.8)$$

Where n is the number of attributes.

### 2.3.4 Decision Tree

A Decision Tree is a part of the Supervised Learning algorithm used for classification problems. It is used for both categorical and continuous dependent variables. The population is divided into two or more similar sets in this algorithm. This method is a graphical representation that makes use of branching methodology. It represents all probable results of a decision based on certain conditions. In a Decision Tree, an Internal node is used to represent a test on the attribute. The outcome of the test is represented by a branch of the tree and a particular class label is represented by the leaf node. All the attributes computed then the decision will take based on that. The classification rules are represented through the path from the root to the leaf node.

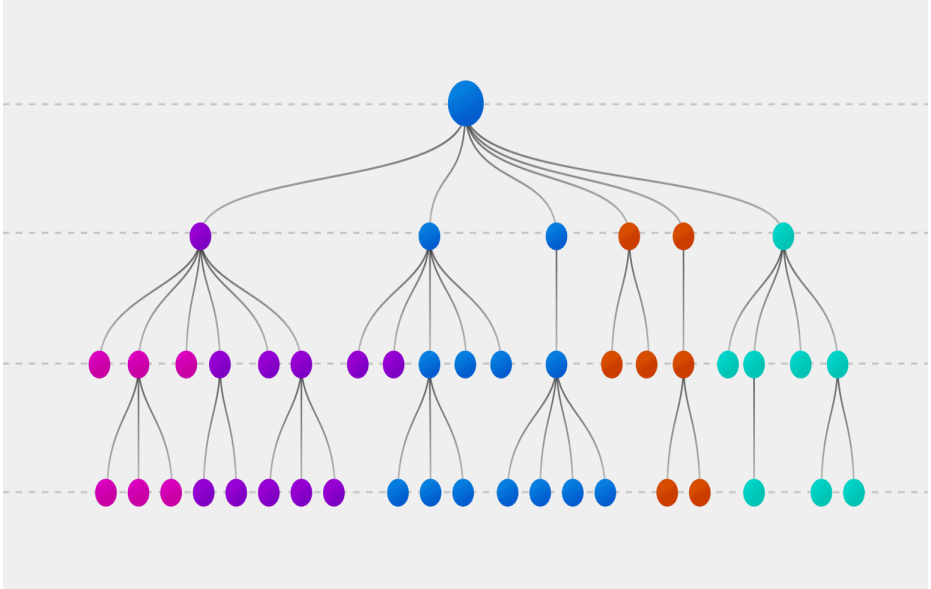

**Figure 2.3** Decision Tree

Source: Explorium, 2019

A Decision Tree (DT) is used for the classification tasks by continuous separating the features in branches into the tree (Sen et al., 2012). Information Gain is used to determine that attribute that splits the features into two subgroups. The root node is the first node; the next nodes are called leaf nodes. This process repeats until the tree is fully built. The node referred to as the end node is shown in Figure 2.3.

It has the advantage that it uses rules that are easily understandable and interpreted by users. It does not require complex data preparation. It performs on both categorical and numerical data.

It uses entropy to define the sample homogeneity. If entropy is zero, it means samples are fully homogeneous, if entropy is one, the samples are equally divided.

$$Info\ (D) = \sum -p_i \log_2 p_i \qquad (2.9)$$

Information Gain is based on entropy after splitting the dataset by an attribute as shown in Equation 2.10.

$$Gain_{(T,x)} = E(T) - E(T,x) \qquad (2.10)$$

### 2.3.5 K-Nearest Neighbor

K-Nearest Neighbor is used for both classification and regression problems. In this algorithm, all available cases are stored and new cases categorized by a majority vote

of their neighbours. The case is most common for the class among its nearest neighbors. The distance from the neighbor classes is measured by a distance function (Figure 2.4).

KNN is a non-probabilistic model for classification study, when there is no prior knowledge about the data preparation. A similarity measure is used for classification. This algorithm picks a value of k which uses several neighbors. For parameter tunning, it begins with k=1, then the algorithm searches for one nearest observation until the best value for k is found.

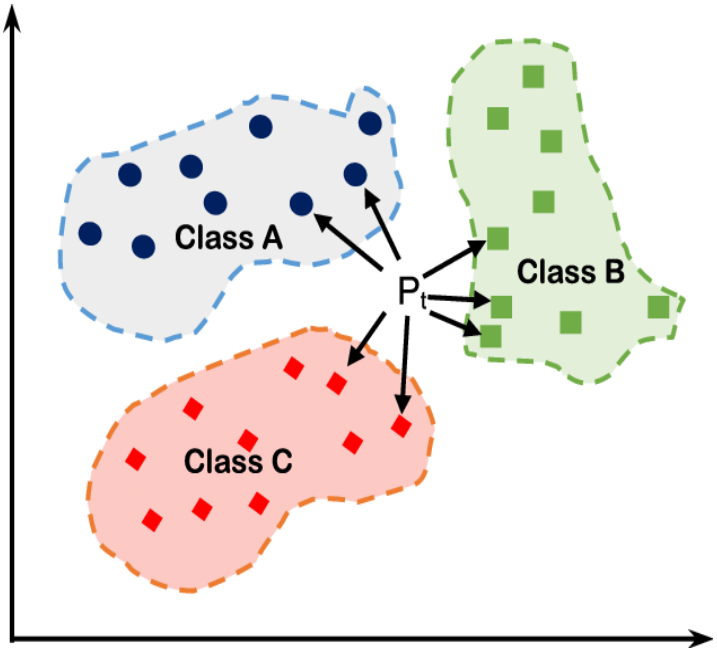


**Figure 2.4** K Nearest Neighbor

Source: Atallah, et al., 2019

The Following steps are followed for KNN implementation:

- Load data
- Initialize the k value
- Calculate distance between target variable and each record of training data using Euclidean distance
- Sort all calculated distance to find the nearest distance
- Find out predicted class

### 2.3.6 Random Forest

Random Forest is an ensemble method that has been used to achieve better performance in many classification problems (Cutter et al., 2007). In this many classifications, trees

are built from the collected dataset. Prediction is performed by combining these trees. It is difficult to interpret, but the combination of these trees gives the high-performance prediction and improved results.

It follows the bagging procedure with some improvements. The bagging procedure is also called bootstrapping. In this procedure, samples are taken from training data repeatedly. The original dataset is divided into different subsets as shown in Figure 2.5. Each of these subsets is used to train a Decision Tree. One subset is used as test data i.e., used to determine the predictive performance of each tree. The final prediction is based on combining all prediction made by each tree based on the voting majority (James et al., 2013).

It is a prominent ensemble method that is used to generate many trees and then combine the results. (Breiman, 2001).

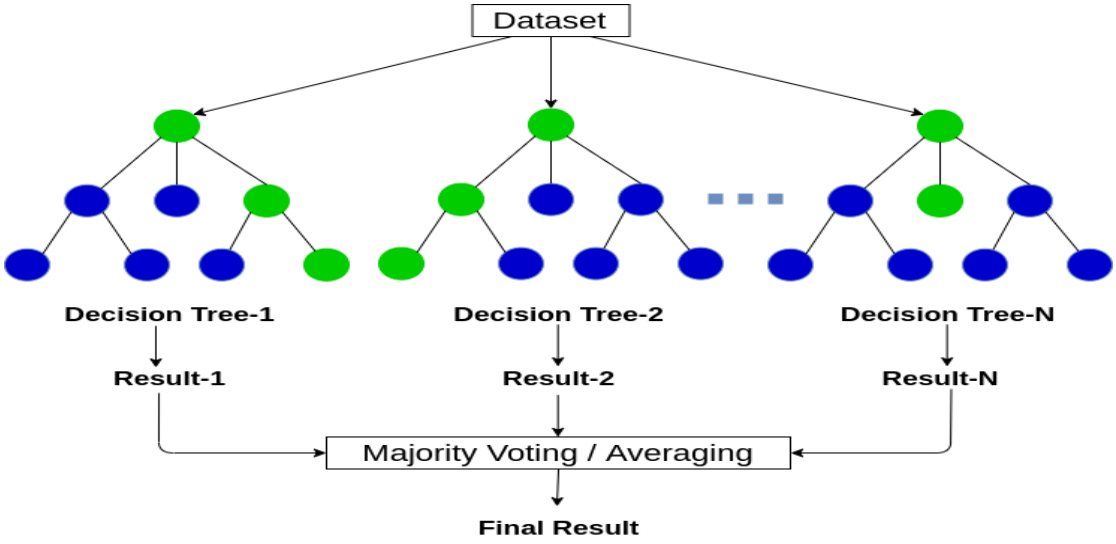


**Figure 2.5** Random Forest

Source: Sharma, 2020

### 2.3.7 Gradient Boosting

Gradient Boosting (GB) is a boosting method based on a different constructive strategy of ensemble formation. The idea behind boosting is to add new models to the ensemble, in sequence. At each repetition, a new weak, base learner model is trained, based on the error of the whole ensemble.

A Gradient decent based formation of boosting methods was derived (Freund & Schapire, 1997; Friedman, 2001). This model's key is learning from previous mistakes. This is a generalization of boosting to the arbitrary differentiable loss function.

$$F_m(x) = F_{m-1}(x) + y_m h_m(x) \qquad (2.11)$$

This algorithm uses the boosting technique, where a number of weak learners is combined to form a strong learner. In this algorithm first a decision Tree is trained. Then the Decision Tree is applied to train the dataset for prediction. The residual of this DT is calculated and saved as the new y. These steps are repeated till the number of trees to train is reached and the final prediction is made (Figure 2.6).

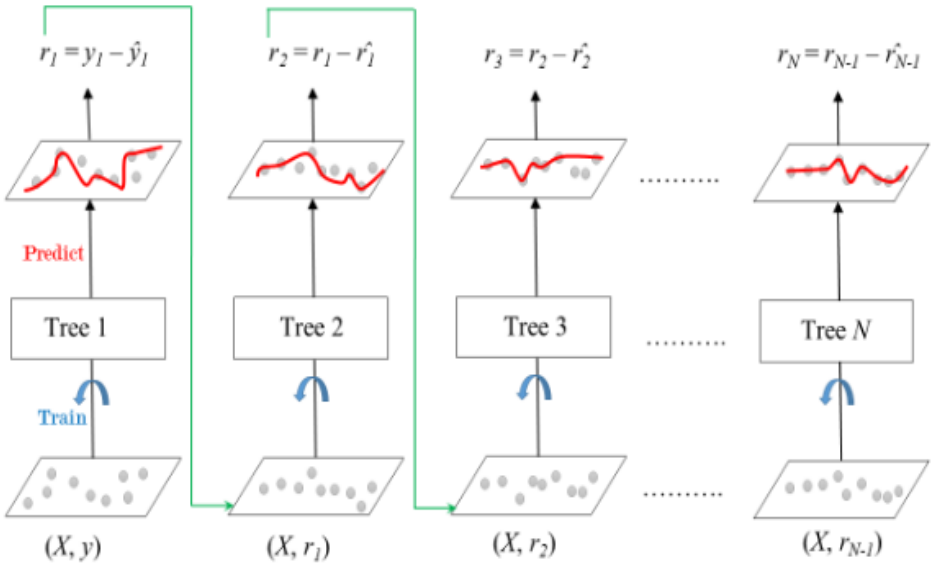


**Figure 2.6** Gradient Boosting

Source: Geeksforgeeks, 2020

### 2.3.8 AdaBoost

AdaBoost or Adaptive boosting method is an ensemble classifier proposed by Yoav Freud and Robert Schapire in 1996. This prooduces a strong classifier by combining multiple classifiers, to increase accuracy. Multiple sequential models are built, each correcting the errors from last models (Avinash Navlani, 2018).

Initially, a training subset is selected. Predictions are made on the whole dataset. Then, predictions and actual values are compared for error calculation. Higher weights are given to the wrongly classified observations so that such observations get a high probability for classification. Also, it assigns weights to the trained classifier in each iteration, based on the accuracy of classifier. This process repeats until the error function does not change or reaches the maximum number of estimators (Figure 2.7).

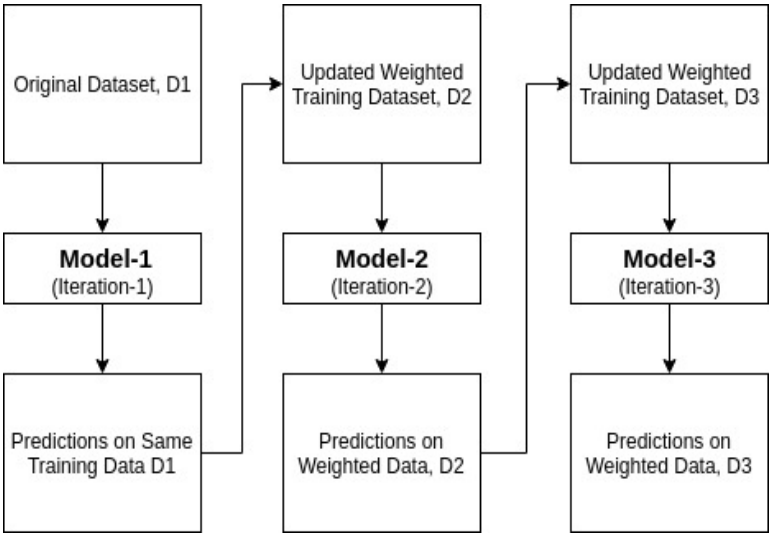


**Figure 2.7** AdaBoost

Source: Navlani A., 2018

### 2.3.9 Bagged Decision Tree

This is an ensemble method that can be used for classification. It is used when the aim is to reduce variance of the Decision Tree. It creates random subsets from the original dataset. This subset includes all the features. A user defined base estimator is fitted on each smaller subset as shown in Figure 2.8. Results of each model are combined to find the final result (Aishwarya Singh, 2018). It improves the accuracy of prediction got by using a single tree, but is difficult to interpret the resulting model.

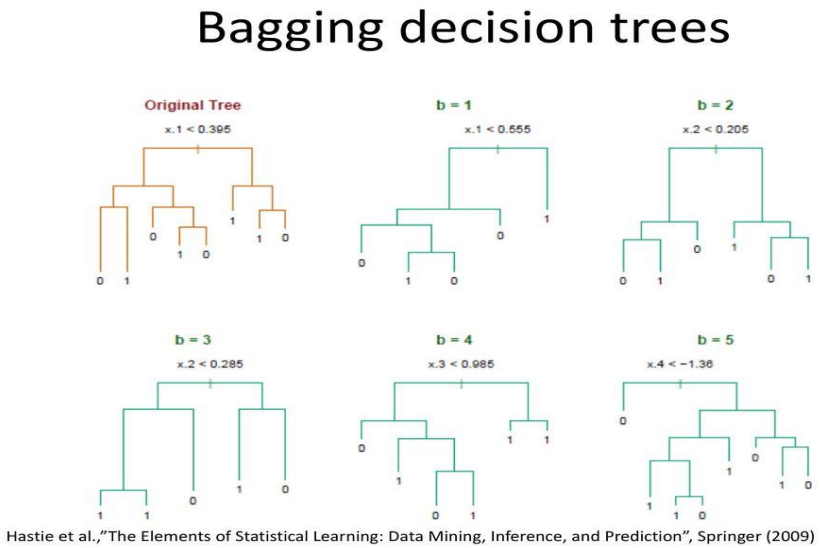


**Figure 2.8** Bagged Decision Tree

Protopapas et al. 2016

## 2.4 Chapter Summary

This chapter introduced the theoretical perspectives of ML techniques and Ensemble Methods. ML classification can be applied to evaluate the performance of students in order to reduce human efforts.

Some selected classification techniques and Ensemble Methods were explained. The techniques were selected based on previous studies in the same field. The selected classifiers are Logistics Regression, Support Vector Machine, K Nearest Neighbor, Decision Tree, Random Forest, Naïve Bayes, Gradient Boosting, AdaBoost and Bagged Decision Tree.

The next chapter will present a Literature Review of different ML techniques used by researchers in the field of higher education, over the previous years. We will discuss separately, first the attributes selected by researchers to evaluate student potential and then the evaluation which was done, based on different ML techniques.