Chapter 2

Big Data Analytics and Algorithms

Alok Kumar, Lakshita Bhargava, and Zameer Fatima

Contents

DOI: 10.1201/9781003175711-2

2.1 Introduction

There is no denying the fact that the digital era is on the horizon, and it is here to stay. In this digital era, a shift is occurring from an industry-based to an information-based economy, which has caused a large amount of data to be accumulated with a mindboggling increase every single day. It is estimated that by 2025 we will be generating 463 exabytes of data every day. This staggering amount of data available is both a boon and a curse for humanity. Improper handling of data can lead to breaches of privacy, an increase in fraud, data loss, and much more. If handled properly, a tremendous growth and enhancement in technology can be achieved. The traditional methods of handling and analyzing data like storing data in traditional relational databases usually perform very poorly in handling big data, the reason being the sheer size of the data. This is where the power of big-data analytics comes into full swing.

The key highlight and main contributions of the chapter include

- The main idea behind writing this chapter is to provide a detailed and structured overview of big-data analytics along with various tools and technology used in the process.
- The chapter provides a clear picture of what big-data analytics is and why it is an extremely important and dominant technology in the current digital era.
- We have also discussed different techniques of big-data analytics along with their relevance in different scenarios.
- A later section of the chapter focuses on some of the most popular and cutting-edge algorithms being used in the process of big-data analytics.
- The chapter concludes with a final section discussing the shortcomings of current data analytics techniques, along with a brief discussion of upcoming technologies that can bridge the gaps present in current techniques.

2.2 Big Data Analytics

Big-data analytics in very simple terms is the process of finding meaningful patterns in a large seemingly unorganized amount of data. The primary goal of big-data analysis is always to provide insights into the source that is responsible for the generation of data. These insights can be extremely valuable for companies to understand the behavior of their customers and how well their product is working in the market. Big-data analytics is also extensively used for revealing product groupings as well as products that are more likely to be purchased together. A mindboggling real-world example of this is the 'diaper-beer' product association found by Walmart upon analyzing its consumer's data. The finding suggested that working men tend to purchase beers for themselves and diapers for their kids together when coming back home from work on Friday night. This led Walmart to put these items together, which saw an increase in the sales of both the items. This finding gives a clear demonstration of the power of big-data analytics for finding product associations, as by using classical product-association techniques it is nearly impossible to find such a bizarre correlation. To get a better understanding of how the process of big-data analytics works in the real world, let's take an example of how an ecommerce company can leverage the power of big-data analytics to increase the sales of their product. In this example, we would consider the broad analysis of two categories of data, data generated by the users in the course of purchasing a product and data generated in after-sales customer service. Big-data analytics techniques like market-basket analysis, customer-product analysis, etc. can be used in the first kind of dataset to find associations like product–product association, customer–product association, or customer–customer association. These findings can be used by the company to improve its product-recommendation system as well as product placement on its portal. Similarly, the results obtained after analysis of after-sales data like customer care phone calls, complaint emails, etc. can be used for training customer-care personnel or even in the development and improvement of smart chatbots. These factors combined can increase the overall customer satisfaction, which can boost the sales number and also help in new-customer acquisition. A surface-level picture of the process is provided in Figure 2.1. Big-data analytics also have found widespread application in the field of medical science. Various data-mining and analytics techniques have been used in a variety of medical applications like disease prediction, genetic programming, patient data management, etc. [1–3]. Data analytics can also be used in educational sectors to analyze students; data and generate better frameworks for enhancing their education [4–5].

2.3 Categories of Big Data Analytics

Big-data analytics is usually classified into four main categories as shown in Figure 2.2. In this section, we will be looking into each of these categories in detail as a separate subsection.

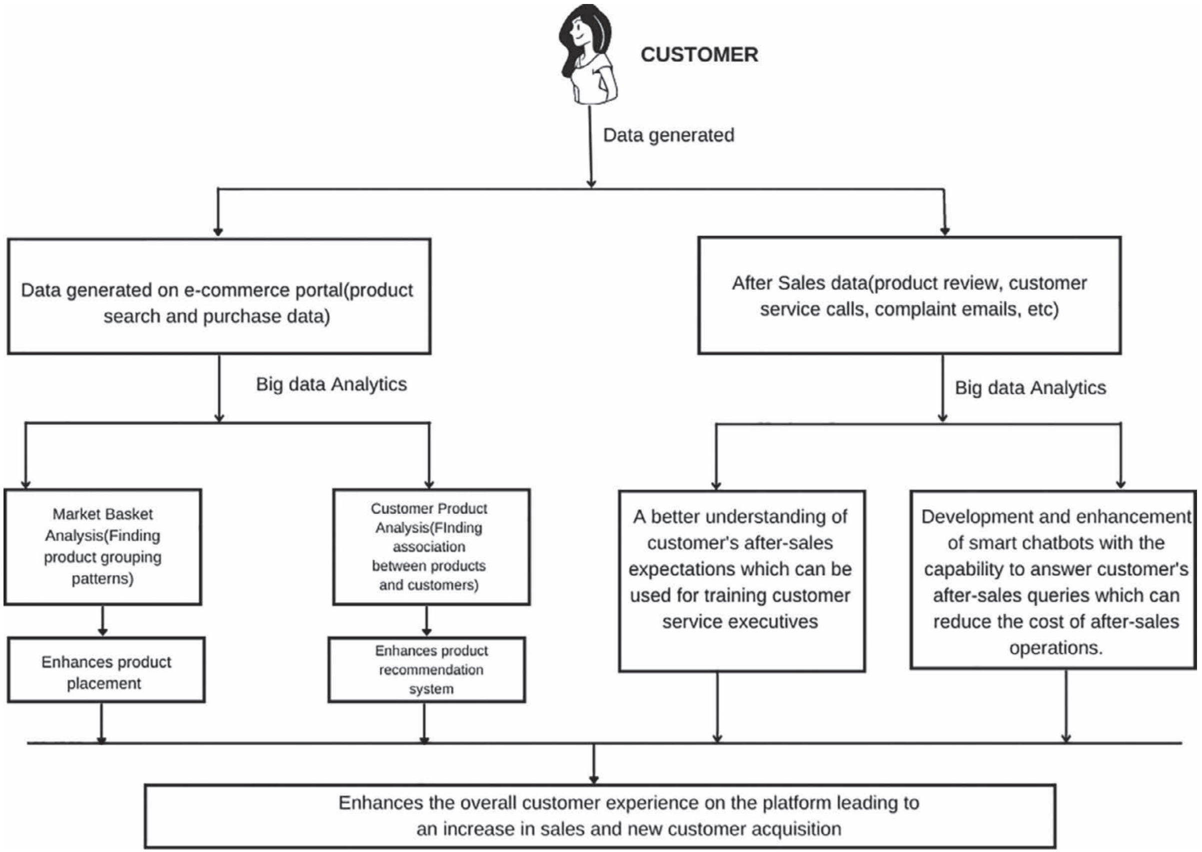
CUSTOMER
Data generated
Data generated on e-commerce portal(product search and purchase data)
After Sales data(product review, customer service calls, complaint emails, etc)
Big data Analytics
Big data Analytics
Market Basket Analysis(Finding product grouping patterns)
Customer Product Analysis(FInding association between products and customers)
A better understanding of customer's after-sales expectations which can be used for training customer service executives
Development and enhancement of smart chatbots with the capability to answer customer's after-sales queries which can reduce the cost of after-sales operations.
Enhances product placement
Enhances product recommendation system
Enhances the overall customer experience on the platform leading to an increase in sales and new customer acquisition

Figure 2.1 Levering Big-Data Analytics in An Ecommerce Company.

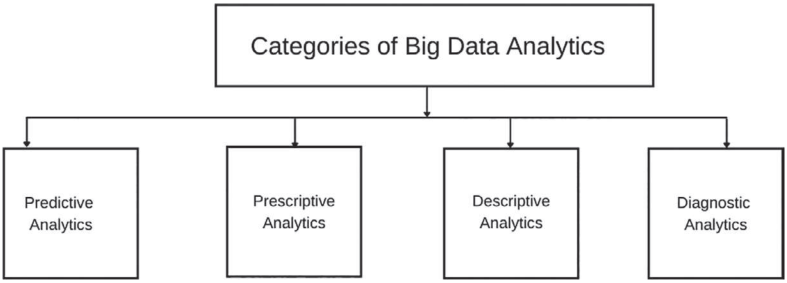

Figure 2.2 Categories of Big-Data Analytics.

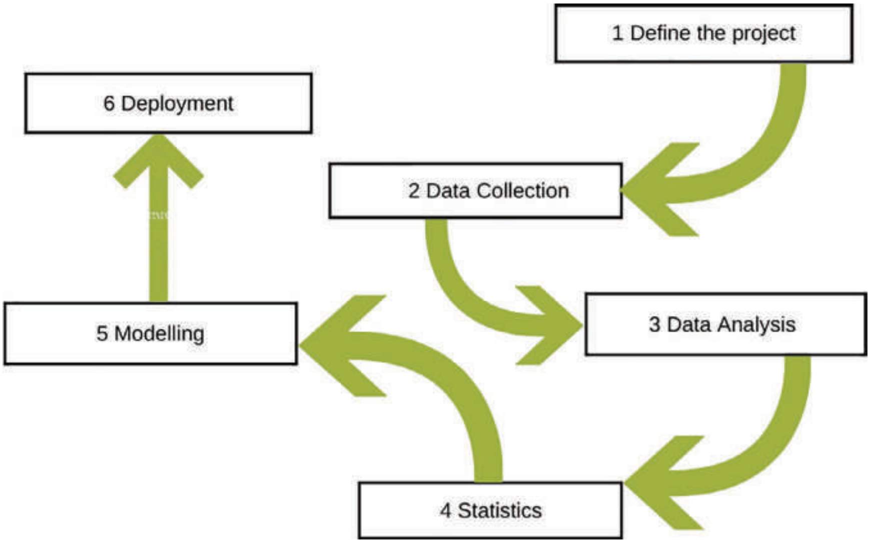

Figure 2.3 Process of Predictive Analytics.

2.3.1 Predictive Analytics

Predictive analytics is a variation of big +-data analytics that is used to make predictions based on the analysis of current data. In predictive analytics, usually historical and transactional data are used to identify risks and opportunities for the future. Predictive analytics empowers organizations in providing a concrete base on which they can plan their future actions. This allows them to make decisions that are more accurate and fruitful compared to the ones taken based on pure assumptions or manual analysis of data. This helps them in becoming proactive and forward-looking organizations. Predictive analytics can even be extended further to include a set of probable decisions that can be made based on the analytics obtained during the process. The whole process of predictive analytics can be broken down into a set of steps as shown in Figure 2.3.

Steps involved in predictive analytics process:

1. Define the project—The first and one of the most important steps in the process of predictive analytics is defining the project. This step consists of identifying different variables like scope and the outcome as well as identifying the dataset on which predictive analytics needs to be executed. This step is extremely crucial as it lays down the foundation for the whole process of data analytics.
2. Data collection—Data is the most fundamental piece of every data-analytics process; it's the same when it comes to predictive analytics. In the data-collection stage organizations collect various types of data through which analytics can take place. The decision to determine the type of data that need to be collected usually depends on the desired outcome of the process established during the project definition stage.
3. Data analysis—The data analysis stage comprises cleaning, transforming, and inspecting data. It is in this stage that patterns, correlations, and useful information about the data are found.
4. Statistics—This is a kind of intermediate stage in which the hypotheses and assumptions behind the model architecture are validated using some existing statistical methods. This step is very crucial as it helps in pointing out any flaws in the logic and highlights inaccuracies that may plague the actual model if unnoticed.
5. Modeling—This stage involves developing the model with the ability to automatically make predictions based on information derived during the data-analytics stage. To improve the accuracy of the model, usually a self-learning module is integrated, which helps in increasing the accuracy of the model over time.
6. Deployment—In the deployment stage, the model is finally deployed on a production-grade server, where it can automatically make decisions and send automated decision reports based on that. It can also be exposed in the form of an application programming interface (API), which can be leveraged by other modules while abstracting the actual complicated logic.
7. Monitoring—Once the deployment is done it is advisable to monitor the model and verify the predictions done by the model on actual results. This could help in enhancing the model and rectifying any minor or major issues that could cripple the performance of the model.

Predictive analytics is being used extensively to tackle a wide variety of problems ranging from simple problems like predicting consumers' behavior on the ecommerce platforms to highly sophisticated ones like predicting the chance of occurrence of a disease in a person based on their medical records. With the advancement in the field of data analytics, the accuracy of predictive analytics models has increased exponentially over the decade, which has enabled their uses in the field of medical science. Maryam et al. have discussed various predictive analytics techniques for predicting

Drug Target Interactions(DTIs) based on analysis of standard datasets [6]. Shakil et al. have proposed a method for predicting dengue disease outbreaks using a predictive analytics tool Weka [1].

2.3.2 Prescriptive Analytics

Prescriptive analytics is a branch of data analytics that helps in determining the best possible course of action that can be taken based on a particular scenario. Prescriptive analytics unlike predictive analytics doesn't predict a direct outcome but rather provides a strategy to find the most optimal solution for a given scenario. Out of all the forms of business analytics, predictive analytics is the most sophisticated type of business analytics and is capable of bringing the highest amount of intelligence and value to businesses [7].

2.3.2.1 How Prescriptive Analytics Works

Prescriptive analytics usually relies on advanced techniques of artificial intelligence, like machine learning and deep learning, to learn and advance from the data it acquires, working as an autonomous system without the requirement of any human intervention. Prescriptive-analytics models also have the capability to adjust their results automatically as new data sets become available.

2.3.2.2 Examples of Prescriptive Analytics

The power of prescriptive analytics can be leveraged by any data-intensive business and government agency. A space agency can use prescriptive analytics to determine whether constructing a new launch site can endanger a species of lizards living nearby. This analysis can help in making the decision to relocate of the particular species to some other location or to change the location of the launch site itself.

2.3.2.3 Benefits of Prescriptive Analytics

Prescriptive analytics is one of the most efficient and powerful tools available in the arsenal of an organization's business intelligence. Prescriptive analytics provides an organization the ability to:

1. Discover the path to success—Prescriptive-analytics models can combine data and operations to provide a road map of what to do and how to do it most efficiently with minimum error.
2. Minimize the time required for planning—The outcome generated by prescriptive-analytics models helps in reducing the time and effort required by the data team of the organization to plan a solution, which enables them to quickly design and deploy an efficient solution

3. Minimize human interventions and errors—Prescriptive-analytics models are usually fully automated and require very few human interventions, which makes them highly reliable and less prone to error compared to the manual analysis done by data scientists.

2.3.3 Descriptive Analytics

Descriptive analytics answers the question of what has happened. The process of descriptive analytics uses a large amount of data to find what has happened in a business for a given period and also how it differs from another comparable period. Descriptive analytics is one of the most basic forms of analytics used by any organization for getting an overview of what has happened in the business. Using descriptive analytics on historic data, decision-makers within the organization can get a complete view of the trend on which they can base their business strategy. It also helps in identifying the strengths and weaknesses lying within an organization. Being an elementary form of analytics technique, it is usually used in conjunction with other advanced techniques like predictive and prescriptive analysis to generate meaningful results.

2.3.4 Diagnostic Analytics

The branch of diagnostic analytics comprises a set of tools and techniques that are used for finding the answer to the question of why certain things happened. Diagnostic analytics takes a deep dive into the data and tries to find valuable hidden insights. Diagnostic analytics is usually the first step in the process of business analytics in an organization. Diagnostic analytics, unlike predictive or prescriptive analytics, doesn't generate any new outcome; rather, it provides the reasoning behind already known results. Techniques like data discovery, data mining, drill-down, etc. are used in the process of diagnostic analytics.

2.3.4.1 Benefits of diagnostic analytics

Diagnostic analytics allows analysists to translate complex data into meaningful visualizations and insights that can be taken advantage of by everyone. Diagnostic analytics also provides insight behind the occurrence of a certain result. This insight can be used to generate predictive- or prescriptive-analytics models.

A comparison of all these four analytics processes along with the critical question answered by each one of them is shown in Table 2.1 and Figure 2.4 respectively.

2.4 Big Data Analytics Algorithms

In the current digital era, data is the new gold. Every organization nowadays understands the importance of having a stockpile of data at its disposal. Companies like Google, Microsoft, and Facebook are dominating the modern era, and a big credit

Table 2.1 Comparison of Different Categories of Data Analytics

Category of classification	*Predictive*	*Prescriptive*	*Descriptive*	*Diagnostic*
Source of data	Uses historical data	Uses historical data	Uses historical data	Uses historical data
Data manipulation	Fills in gaps in available data	Estimates outcomes based on variables	Reconfigures data into easy-to-read format	Identifies anomalies
Role of analytics	Creates data models	Offers suggestions about outcomes	Describes the state of business operation	Highlights data trends
Technique used	Forecasts potential future outcomes	Uses algorithms, machine learning, and AI	Learns from the past	Investigates underlying issues
Critical question answered	Answers 'What might happen?'	Answers '*If, then* questions'	Answer '*What* questions'	Answer '*Why* questions'

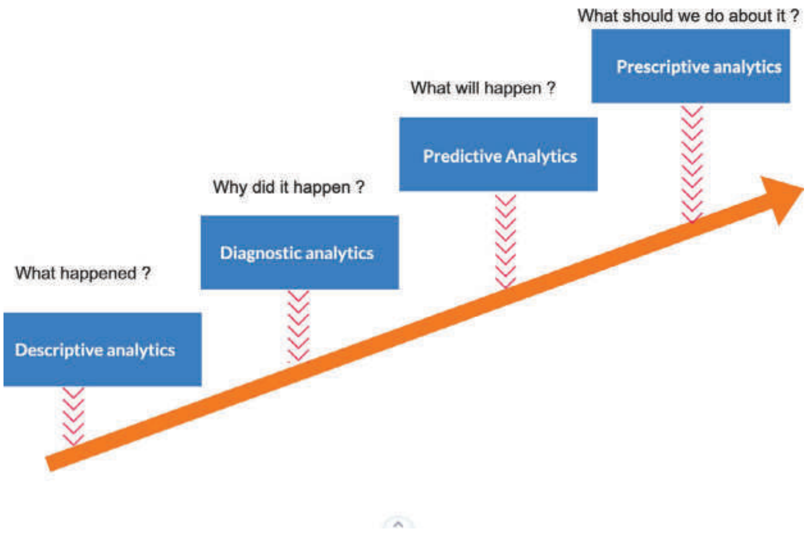

Figure 2.4 Critical Questions Answered by Different Analytics Techniques.

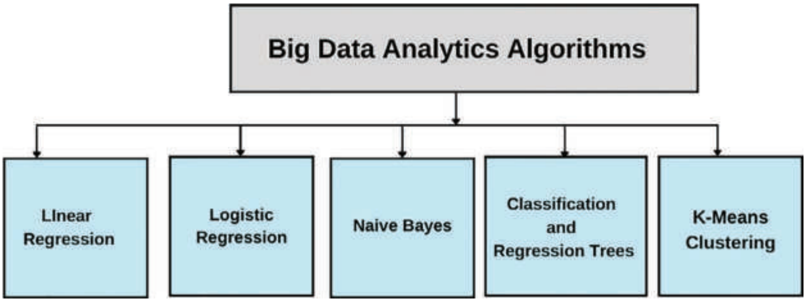

Figure 2.5 Big-Data Analytics Algorithms.

for that goes to the mammoth data stores they have at their disposal. Having such huge data stores at their disposal has enabled these companies to push the boundaries of technological advancement in a way that was never seen before. A burning example that exhibits the power of data and what can be achieved through its proper analytics is Google Maps. Built on top of data pipelines containing a huge amount of dynamic and diverse data collected by Google from multiple sources, it is a piece of technology that seems like something straight from the future.

But having data alone is not sufficient. Data on its own is useless and becomes meaningful only when proper analysis of that data is done. With an unprecedented increase in the amount of data generated in the last couple of years, it has become more necessary now than ever to have fast and efficient data-analytics algorithms at our disposal as the classical methods of data analysis using graphs or charts are simply not enough to keep up with this huge amount of data otherwise also known as Big Data. To solve this problem, data scientists all over the world have developed and are in the process of developing new advanced algorithms for analyzing big data efficiently. To discuss all of these algorithms is beyond the scope of this chapter, hence we will keep our focus on the five most popular big-data analytics algorithms that usually form the basis of the majority of high-performance analytics models. These algorithms are shown in Figure 2.5 and discussed afterward.

2.4.1 Linear Regression

Linear regression is a kind of statistical test performed on a dataset to define and find the relation between considered variables [8]. Linear regression is one of the most popular and frequently used statistical analysis algorithms. Being a very simple yet extremely powerful algorithm for data analysis, it is used by data scientists extensively for designing simple as well as complicated analytical models.

Linear regression, as the name suggests, is a simple linear equation that combines the input values (x) and then generates the solution as a predicted output (y). In the linear-regression model, a scale factor is assigned to each of the input values or independent variables, which is also known as a coefficient and is symbolized using the Greek letter Beta (β). An extra coefficient, also known as intercept or bias coefficient, is added to the equation, which provides an additional degree of freedom to the line. If the linear-regression equation contains a single dependent variable (y) and a single independent variable (x), it is known as *univariate regression* and is represented by equation 2–1:

$$y = \beta_1 * x + \beta_0 \tag{2–1}$$

y = *dependent variable*
x = *independent variable*
β_1 = *scale factor*
β_0 = *bias coefficient*

The regression model with more than one independent variable is known as *multivariate regression*. In a multivariate-regression model, an attempt is made to account for the variation of independent variables in the dependent variable synchronically [9]. The equation of multivariate regression is an extension of univariate regression and is represented in equation 2–2:

$$y = \beta_0 + \beta_1 * x_1 + \cdots + \beta_n * x_n + \varepsilon \tag{2–2}$$

y = *dependent variable*
x = *independent variable*
$(\beta_1 - \beta_n)$ = *scale factor*
β_0 = *bias coefficient*
ε = *error*

2.4.1.1 Preparing a Linear-Regression Model

Preparing a linear-regression model, also known as model training, is the process of estimating the coefficients of the equation to find the best-fitting line for our dataset. There are several methods for training a linear-regression model. In this section, we will be discussing three of the most commonly used methods among them.

1. Simple Linear Regression—Simple linear regression is a technique for training linear-regression models when there is only one input—or, better to say, only one independent variable—in the equation. In the method of simple linear regression, model statistical properties from the data like mean, standard deviation, correlations, and covariance are calculated, which are used for estimating the coefficients and hence finding the best-fitting line.

2. Least Square—The method of least square is used when there are multiple dependent variables and an estimation of the values of the coefficients is required. This procedure seeks to attenuate the sum of the squared residuals. The method suggests that, for a given regression curve, we can calculate the space from each datum to the regression curve, square it, and determine the sum of all of the squared errors together. This is often the value that the method of least squares needs to attenuate.
3. Gradient descent—The method of gradient descent is used in the scenario when there are one or more inputs and there is a requirement for optimizing the value of the coefficient, which is done by an iterative minimization of the error of the model on training data. The algorithm starts by assigning random values to every coefficient. Calculating the sum of squared errors for all pairs of input and output values is the next step in the process of gradient descent. A learning rate is associated, which acts as a multiplier with which the value of coefficients are updated with the goal of minimizing the error. This process gets terminated when either minimum-squared sum has been achieved or any further improvement is not feasible.

   The variation of gradient descent using a rectilinear-regression model is more commonly used as it is relatively straightforward to understand. This algorithm finds application in the scenario when the dataset is large and hence won't fit into the memory.

2.4.1.2 Applications of Linear Regression

Linear regression is a simple yet very sophisticated algorithm that finds application in a wide variety of fields. Roy et al. have proposed a *Lasso Linear Regression Model* for stock-market forecasting [9]. Zameer et al. have used a linear-regression-based model for predicting crude-oil consumption [10]. In general, linear-regression models are quite good in performing predictive data analytics.

2.4.2 Logistic Regression

The technique of *logistic regression* in big data analytics is used when the variable to be considered is dichotomous (binary). The basis of logistic regression, just like all other regression, is a predictive analysis. Logistic regression is employed to elucidate data and to explain the connection between one dependent binary variable and one or more nominal, ordinal, interval, or ratio-level independent variables.

Logistic regression works on the concept of logit—the natural logarithms of an odds ratio [11]. This type of regression model works quite well when the dependent variable is categorical. Some examples of real-world problems where the dependent variable can be categorical are predicting if the email is spam (1) or not (0) or if a tumor is malignant (1) or safe (0). Logistic regression is a component of a bigger class of algorithms referred to as the generalized linear model (GLM). In 1972,

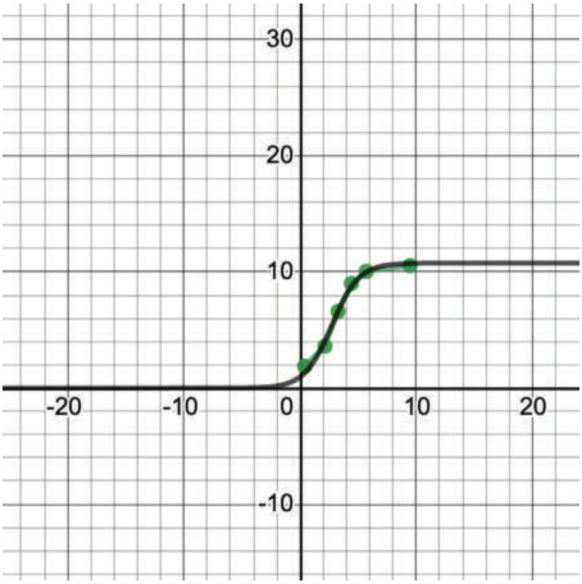

Figure 2.6 A Sample Logistic-Regression Plot.

Nelder and Wedderburn proposed this model in an attempt to supply a way of using rectilinear regression with the issues that weren't directly fitted to the application of rectilinear regression. They proposed a category of various models (linear regression, ANOVA, Poisson regression, etc.), including logistic regression as a special case. Equation 2–3 represents a general equation of logistic regression.

$$loglog\{1-p\} = \beta_0 + \beta_1 * x \tag{2–3}$$

(p/1-p) = odd ratio
x = independent variable
β_1 = *scale factor*
β_0 = *bias coefficient*

In this equation $\{1-p\}$ is the odds ratio. The positive log of an odds ratio usually translates into a probability of success greater than 50%. A sample plot of logistic regression is shown in Figure 2.6.

2.4.2.1 Types of Logistic Regression

1. **Binary Logistic Regression**
   In binary logistic regression, a categorical response can only have two possible outcomes. Example: Spam or Not email.

2. **Multinomial Logistic Regression**
   In multinomial logistic regression, dependent (target) variables can have three or more categories without ordering. Example: predicting which food is preferred more (Veg, Non-Veg, Vegan).
3. **Ordinal Logistic Regression**
   Ordinal logistic regression is a subset of multinomial logistic regression in which dependent (target) variables can have three or more categories but in a defined order. Example: movie rating from 1–5.

2.4.2.2 Applications of Logistic Regression

Logistic regression is a simple yet efficient algorithm that finds application in a wide variety of fields. Due to its predictive nature, logistic regression finds application in fields ranging from education to healthcare. Ramosaco et al. have developed a logistic-regression-based model to study students' performance levels [12]. Alzen et al. have proposed another logistic-regression-based model to find the relationship between the learning assistant model and failure rates in introductory STEM courses [13].

Although linear regression and logistic regression are both regression-based models, they do share a lot of differences. These differences are shown in Table 2.2.

Table 2.2 Difference between Linear and Logistic Regression

Linear Regression	*Logistic Regression*
Linear regression is used to predict the continuous dependent variable using a given set of independent variables.	Logistic regression is used to predict the categorical dependent variable using a given set of independent variables.
Linear regression is used for solving the regression problem.	Logistic regression is used for solving classification problems.
In linear regression, we predict the value of continuous variables.	In logistic regression, we predict the values of categorical variables.
In linear regression, we find the best-fitting line, by which we can easily predict the output.	In logistic regression, we find the S-curve by which we can classify the samples.
The least-square estimation method is used for the estimation of accuracy.	The maximum-likelihood estimation method is used for the estimation of accuracy.
The output of linear regression must be a continuous value, such as price, age, etc.	The output of logistic regression must be a categorical value such as 0 or 1, Yes or No, etc.

Linear Regression	*Logistic Regression*
In linear regression, it is required that the relationship between the dependent variable and independent variable be linear.	In logistic regression, it is not required to have the linear relationship between the dependent and independent variable.
In linear regression, there may be collinearity between the independent variables.	In logistic regression, there should not be collinearity between the independent variables.

2.4.3 Naive Bayes Classifiers

Naive Bayes classifiers are a set of classification algorithms supported by Bayes' theorem. It's not one algorithm but a family of algorithms where all of them share a standard principle, i.e. every pair of features being classified is independent of every other.

Naive Bayes uses the probabilistic approach for constructing classifiers. These classifiers can simplify learning by assuming that features are independent of given class [14]. Naive Bayes classification is a subset of Bayesian decision theory. It's called *naive* because the formulation makes some naive assumptions [15].

The main assumption that Naive Bayes classifiers make is that the value of a specific feature is independent of the value of the other feature. Despite having an oversimplified assumption, Naive Bayes classifiers tend to perform well even in complex real-world scenarios. The main advantage that Naive Bayes classifiers have over other classification algorithms is the requirement of a little amount of training data for estimating the parameters necessary for classification, which is used for an incremental training of the classifier.

2.4.3.1 Equation of the Naive Bayes Classifiers

To understand the equation of Naive Bayes classifiers we need to understand Bayes' theorem, which is the fundamental theorem on which Naive Bayes classifiers work.

Bayes' theorem

Bayes' theorem finds the probability of the occurrence of an event, given the probability of another event that has already occurred. Bayes theorem is stated mathematically as shown in equation 2–4:

$$P\left(\frac{A}{B}\right) = \frac{P\left(\frac{B}{A}\right) * P(A)}{P(B)} \tag{2–4}$$

P(A) = Probability of occurrence of event A
P(B) = Probability of occurrence of event B
P(A/B) = Probability of A given B
P(B/A) = Probability of B given A

Bayes' theorem can be extended to find equations of various Naive Bayes classifiers.

2.4.3.2 Application of Naive Bayes Classifiers

Naive Bayes classifiers, despite having certain limitations and assumptions, work quite well for solving classification problems. Karthika and Sairam propose a classification methodology utilizing the Naive Bayesian classification algorithm for the classification of persons into different classes based on various attributes representing their educational qualification [16]. Qin et al. research classifying multilabel data based on Naive Bayes classifiers, which can be extended to multilabel learning [17].

2.4.4 Classification and Regression Trees

Classification and regression trees (CART) is a term coined by Leo Breiman to allude to the decision tree class of algorithms that are used to solve the classification and regression predictive analytics problems.

Traditionally, this calculation is alluded to as 'decision trees'; however, in certain programming languages like R they are alluded to by the more present-day term CART. The CART algorithms give an establishment for some other significant algorithms like bagged decision-tree algorithms, random-forest algorithms, and boosted decision-tree algorithms.

2.4.4.1 Representation of CART Model

The CART model can be represented as a binary tree. Each node in the tree represents a single input variable (x) and a split point theorem variable, and the leaf node is represented using an output variable (y), which is utilized for forecasting.

For example, suppose a dataset having two input variables (x) of height in centimeter and weight of a person in kilogram the output variable (y) will tell whether the sex of the person is male or female. Figure 2.7 represents a very simple binary decision tree model.

A straightforward way for making predictions using the CART model is with the help of its binary tree representation. The traversal of the tree starts with the evaluation of a specific input starting with the root node of the tree. Each input variable in the CART model can be thought of as a dimension in an n-dimensional space. The decision tree in this model splits this plane into rectangles for two input

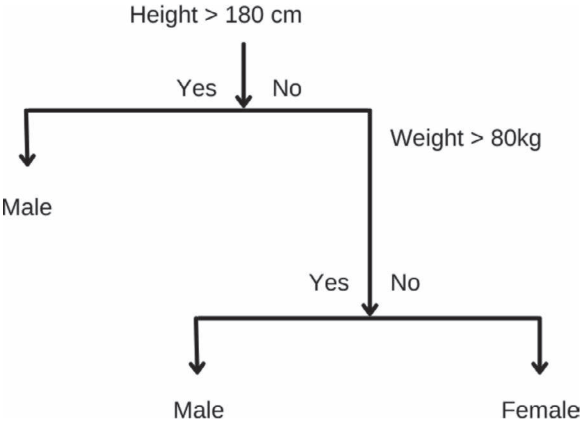

Figure 2.7 Representation of Binary Decision-Tree Model.

variables or into hyperrectangles for higher inputs. The input data gets filtered through the tree and gets placed in one of the rectangles, whereas the prediction made by the model is the output value for the same rectangle; this gives us some idea about the type of decisions that a CART model is capable of making, e.g. boxy decision boundaries.

2.4.4.2 Application of Classification and Regression Trees

Pham et al. have used a classification and regression tree-based model for predicting the rainfall-induced shallow landslides in the state of India based on a dataset of 430 historic landslide locations [18]. Pouliakis et al. have done a study on CART-based models to estimate the risk for cervical intraepithelial neoplasia [19]. Iliev et al. have proposed a CART-based model for modeling the laser output power of a copper bromide vapor laser [20].

2.4.5 K-Means Clustering

K-means clustering is a very simple yet popular data-analytics algorithm. It is an unsupervised algorithm as it capable of drawing conclusions from datasets having only input variables without the requirement of having known or labeled outcomes. The goal of the K-means algorithm is very basic: just group similar data points and reveal the pattern present in the dataset. K-means tries to find a predefined number (k) of the cluster in the dataset. A cluster in very simple terms can be thought of as a group of similar data points. The prerequisite of the algorithm is the target number *k*, which denotes the number of centroids required by us. A centroid can either be a real or an imaginary point that represents the center of one single cluster. Each information point is designated for every one of the groups by reducing the in-cluster sum of squares. The K-means algorithm distinguishes the predefined number of centroids and afterward allots each data point to the nearest cluster, with the goal being to keep the centroids as tiny as could be expected. The 'means' in the K-means alludes to the aggregation of the information or, say, finding the centroid.

2.4.5.1 How K-Means Clustering Works

For handling the learning information, the K-means algorithm in data analytics begins with a set of randomly selected centroids; these are utilized as the starting point for each cluster and afterward perform iterative calculations to improve the places of the centroids.

It stops making and optimizing cluster when either of the conditions is met:

- The centroids have stabilized and the algorithm can proceed further, i.e. the clustering has been successful.
- The predefined number of iterations has been reached.

2.4.5.2 The K-Means Clustering Algorithm

The K-means clustering algorithm follows the approach of expectation-maximization. The expectation step is assigning the data point to the closet cluster. The maximization step is finding the centroid of each of these clusters. The final goal of the K-means algorithm is to minimize the value of squared error function given as:

$$J(V)=\sum_{i=1}^{c}\sum_{j=1}^{c_i}\left(\left|x_i - v_j\right|\right)^2$$

$\left|x_i - v_j\right|$ is the Euclidean distance between x_i and v_j

2.4.5.3 Application of K-Means Clustering Algorithms

Being a high performing, unsupervised learning algorithm, K-means finds application in a wide variety of fields. Due to its popularity, researchers have created different hybrid versions of this algorithm that are being used extensively in numerous fields. Youguo & Haiyan have developed a clustering algorithm on top of K-means clustering, which provides greater dependence to choose the initial focal point [21]. Shakil and Alam have devised a method for data management in the cloud-based environment on the basis of the K-means clustering algorithm [22]. Alam and Kishwar have categorized various clustering techniques that have been applied to web search results [23]. Alam and Kishwar have proposed an algorithm for web-search clustering based on K-means and a heuristic search [24].

2.5 Conclusion and Future Scope

In this chapter, we looked into the basics of data analytics along with its application in the real world. We also looked into various categories of data analytics along with some of the most commonly used data-analytics algorithms as well as their applications to the real-world scenario. Apart from the algorithms discussed in this chapter, data scientists all over the world have been working on designing faster and more efficient algorithms. The idea of using neural-network-based algorithms has been also proposed by data scientists [25, 32]. With the rise of quantum computing in the last couple of years, scientists are also looking forward to the possibility of leveraging the power of quantum computers in big-data analytics [26]. Cloud-based big-data analytics is also becoming quite popular as it can leverage the power of cloud computing for big-data analytics [27–31]. With these new technological advancements on the horizon, it can be safely assumed that the future of big-data analytics is going to be bright and exciting.

References

1. Shakil, K. A., Anis, S., & Alam, M. (2015). Dengue disease prediction using weka data mining tool. *arXiv preprint arXiv:1502.05167*.
2. Khan, M. W., & Alam, M. (2012). A survey of application: Genomics and genetic programming, a new frontier. *Genomics*, *100*(2), 65–71.
3. Shakil, K. A., Zareen, F. J., Alam, M., & Jabin, S. (2020). BAMHealthCloud: A biometric authentication and data management system for healthcare data in cloud. *Journal of King Saud University-Computer and Information Sciences*, *32*(1), 57–64.
4. Khanna, L., Singh, S. N., & Alam, M. (2016, August). Educational data mining and its role in determining factors affecting students academic performance: A systematic review. In *2016 1st India International Conference on Information Processing (IICIP)* (pp. 1–7). New York: IEEE.
5. Shakil, K. A., Sethi, S., & Alam, M. (2015, March). An effective framework for managing university data using a cloud based environment. In *2015 2nd International Conference on Computing for Sustainable Global Development (INDIACom)* (pp. 1262–1266). New York: IEEE.
6. Bagherian, M., Sabeti, E., Wang, K., Sartor, M. A., Nikolovska-Coleska, Z., & Najarian, Z. (2021). Machine learning approaches and databases for prediction of drug—target interaction: a survey paper. *Briefings in Bioinformatics*, *22*(1), 247–269. https://doi.org/10.1093/bib/bbz157
7. Šikšnys, L., Pedersen, T. B., Liu, L., & Özsu, M. (2016). Prescriptive analytics. *Encyclopedia of Database Systems*, 1–2.
8. Kaya Uyanık, G., & Güler, N. (2013). A study on multiple linear regression analysis. *Procedia—Social and Behavioral Sciences*, *106*, 234–240. https://doi.org/10.1016/j.sbspro.2013.12.027.
9. Roy, S. S., Mittal, D., Basu, A., & Abraham, A. (2015). Stock market forecasting using LASSO linear regression model. In: Abraham, A., Krömer, P., &

Snasel, V. (eds.), *Afro-European Conference for Industrial Advancement. Advances in Intelligent Systems and Computing*, vol. 334. Cham: Springer. https://doi.org/10.1007/978-3-319-13572-4_31.

10. Fatima, Z., Kumar, A., Bhargava, L., & Saxena, A. (2019). Crude oil consumption forecasting using classical and machine learning methods. *International Journal of Knowledge Based Computer Systems*, *7*(1), 10–18.
11. Peng, J., Lee, K., & Ingersoll, G. (2002). An introduction to logistic regression analysis and reporting. *Journal of Educational Research*, *96*, 3–14. https://doi.org/10.1080/00220670209598786.
12. Ramosaco, M., Hasani, V., & Dumi, A. (2015). Application of logistic regression in the study of students' performance level (Case Study of Vlora University). *Journal of Educational and Social Research*, *5*(3). https://doi.org/10.5901/jesr.2015.v5n3p239.
13. Alzen, J. L., Langdon, L. S., & Otero, V. K. (2018). A logistic regression investigation of the relationship between the Learning Assistant model and failure rates in introductory STEM courses. *International Journal of STEM Education*, *5*, Article number 56. https://doi.org/10.1186/s40594-018-0152-1.
14. Rish, I. (2001). An empirical study of the naïve bayes classifier. *IJCAI 2001 Work Empir Methods Artif Intell*, *3*.
15. Kaviani, P., & Dhotre, S. (2017). Short survey on naive bayes algorithm. *International Journal of Advance Research in Computer Science and Management*, *4*.
16. Karthika, S., & Sairam, N. (2015). A naïve bayesian classifier for educational qualification. *Indian Journal of Science and Technology*, *8*. https://doi.org/10.17485/ijst/2015/v8i16/62055.
17. Qin, F., Tang, X., & Cheng, Z. (2012). Application and research of multi_label Naïve Bayes Classifier. In *Proceedings of the 10th World Congress on Intelligent Control and Automation* (pp. 764–768). New York: IEEE. https://doi.org/10.1109/WCICA.2012.6357980.
18. Pham, B. T., Tien Bui, D., & Prakash, I. (2018). Application of classification and regression trees for spatial prediction of rainfall-induced shallow landslides in the Uttarakhand Area (India) using GIS. In: Mal, S., Singh, R., & Huggel, C. (eds.), *Climate Change, Extreme Events and Disaster Risk Reduction. Sustainable Development Goals Series*. Cham: Springer. https://doi.org/10.1007/978-3-319-56469-2_11.
19. Pouliakis, A., Karakitsou, E., Chrelias, C., Pappas, A., Panayiotides, I., Valasoulis, G., Kyrgiou, M., Paraskevaidis, E., Karakitsos, P. (2015). The application of classification and regression trees for the triage of women for referral to colposcopy and the estimation of risk for cervical intraepithelial neoplasia: A study based on 1625 cases with incomplete data from molecular tests. *BioMed Research International*, *2015*, Article ID 914740, 10 p. https://doi.org/10.1155/2015/914740.
20. Iliev, I. P., Voynikova, D. S., & Gocheva-Ilieva, S. G. (2013). Application of the classification and regression trees for modeling the laser output power of a copper bromide vapor laser. *Mathematical Problems in Engineering*, *2013*, Article ID 654845, 10 p. https://doi.org/10.1155/2013/654845.
21. Li, Y., & Wu, H. (2012). A clustering method based on K-means algorithm. *Physics Procedia*, *25*, 1104–1109. https://doi.org/10.1016/j.phpro.2012.03.206.
22. Shakil, K. A., & Alam, M. (2014). Data management in cloud based environment using k-median clustering technique. *International Journal of Computer Applications*, *3*, 8–13.

23. Alam, M., & Sadaf, K. (2013). A review on clustering of web search result. In: *Advances in Computing and Information Technology* (pp. 153–159). Berlin, Heidelberg: Springer.
24. Alam, M., & Sadaf, K. (2015). Web search result clustering based on heuristic search and K-means. *arXiv preprint arXiv:1508.02552*.
25. Mamatha, C., Reddy, P., Kumar, M. A., & Kumar, S. (2017). Analysis of big data with neural network. *International Journal of Civil Engineering and Technology*, *8*, 211–215.
26. Shaikh, T. (2016). Quantum computing in big data analytics: A survey. *Conference: 2016 IEEE International Conference on Computer and Information Technology (CIT)* (pp. 112–115). https://doi.org/10.1109/CIT.2016.79.
27. Khan, S., Shakil, K. A., & Alam, M. (2017). Cloud based big data analytics: A survey of current research and future directions. *Big Data Analytics*, Print ISBN: 978-981-10-6619-1, Electronic ISBN: 978-981-10-6620-7, (pp. 629–640). Springer.
28. Alam, M. (2012). Cloud algebra for cloud database management system. *The Second International Conference on Computational Science, Engineering and Information Technology (CCSEIT-2012)*, October 26–28, Coimbatore, India, Proceeding published by ACM.
29. Alam, M. (2012). Cloud algebra for handling unstructured data in cloud database management system. *International Journal on Cloud Computing: Services and Architecture (IJCCSA)*, *2*(6), ISSN: 2231–5853 [Online]; 2231–6663 [Print]. https://doi.org/10.5121/ijccsa.2012.2603, Taiwan.
30. Alam, M., & Shakil, K. (2013). Cloud database management system architecture. *UACEE International Journal of Computer Science and its Applications([ISSN 2250–3765)*, *3*, 27–31.
31. Alam, B., Doja, M. N., Alam, M., & Malhotra, S. (2013). 5-layered architecture of cloud database management system. *AASRI Procedia Journal*, *5*, 194–199, ISSN: 2212–6716, Elsevier.
32. Alam, M., Shakil, K. A., Mohd. Javed, S., & Ambreen, M. A. (2014). Detect and filter traffic attack through cloud trace back and neural network. *The 2014 International Conference of Data Mining and Knowledge Engineering (ICDMKE), Imperial College, London, UK, 2–4 July*. Hong Kong: IAENG.