Noname manuscript No.
(will be inserted by the editor)

Data Science and Analytics: An Overview from Data-Driven Smart Computing, Decision-Making and Applications Perspective

Iqbal H. Sarker[1,2*]

the date of receipt and acceptance should be inserted later

Abstract The digital world has a wealth of data, such as Internet of Things (IoT) data, business data, health data, mobile data, urban data, security data, and many more, in the current age of the Fourth Industrial Revolution (Industry 4.0 or 4IR). Extracting knowledge or useful insights from these data can be used for *smart decision-making* in various applications domains. In the area of data science, *advanced analytics methods* including *machine learning* modeling can provide actionable insights or deeper knowledge about data, which makes the computing process automatic and smart. In this paper, we present a comprehensive view on *"Data Science"* including various types of advanced analytics methods that can be applied to enhance the intelligence and capabilities of an application through smart decision-making in different scenarios. We also discuss and summarize ten potential *real-world application domains* including business, healthcare, cybersecurity, urban and rural data science, and so on by taking into account data-driven smart computing and decision making. Based on this, we finally highlight the challenges and potential *research directions* within the scope of our study. Overall, this paper aims to serve as a reference point on *data science and advanced analytics* to the researchers and decision-makers as well as application developers, particularly from the data-driven solution point of view for real-world problems.

Keywords data science, advanced analytics, machine learning; deep learning; smart computing; decision-making; predictive analytics; data science applications;

1 Introduction

We are living in the age of "data science and advanced analytics", where almost everything in our daily lives is digitally recorded as data [19]. Thus the current electronic world is a wealth of various kinds of data, such as business data, financial data, healthcare data, multimedia data, Internet of Things (IoT) data, cybersecurity data, social media data, etc [112]. The data can be structured, semi-structured, or unstructured, which increases day by day [105]. Data science is typically a "concept to unify statistics, data analysis, and their related methods" to understand and analyze the actual phenomena with data. According to Cao et al. [19] "data science is the science of data" or "data science is the study of data", where a data product is a data deliverable, or data-enabled or guided, which can be a discovery, prediction, service, suggestion, insight into decision-making, thought, model, paradigm, tool, or system. The popularity of "Data science" is increasing day-by-day, which is shown in Fig. 1 according to Google Trends data over the last five years [3]. In addition to data science, we have also shown the popularity trends of the relevant topics such as "Data analytics", "Data mining", "Big data", "Machine learning" in the figure. According to Fig. 1, the popularity indication values for these data-driven domains, particularly "Data science", and "Machine learning" are increasing day-by-day. This statistical information and the applicability of the data-driven smart decision-making in various real-world application areas, motivate us to study briefly on *"Data science"* and machine-learning-based *"Advanced analytics"* in this paper.

Usually, data science is the field of applying advanced analytics methods and scientific concepts to derive useful business information from data. The empha-

[1] Swinburne University of Technology, Melbourne, VIC 3122, Australia.
[2] Department of Computer Science and Engineering, Chittagong University of Engineering & Technology, Bangladesh.
[*] Correspondence: msarker@swin.edu.au (Iqbal H. Sarker); ORCID iD: https://orcid.org/0000-0003-1740-5517

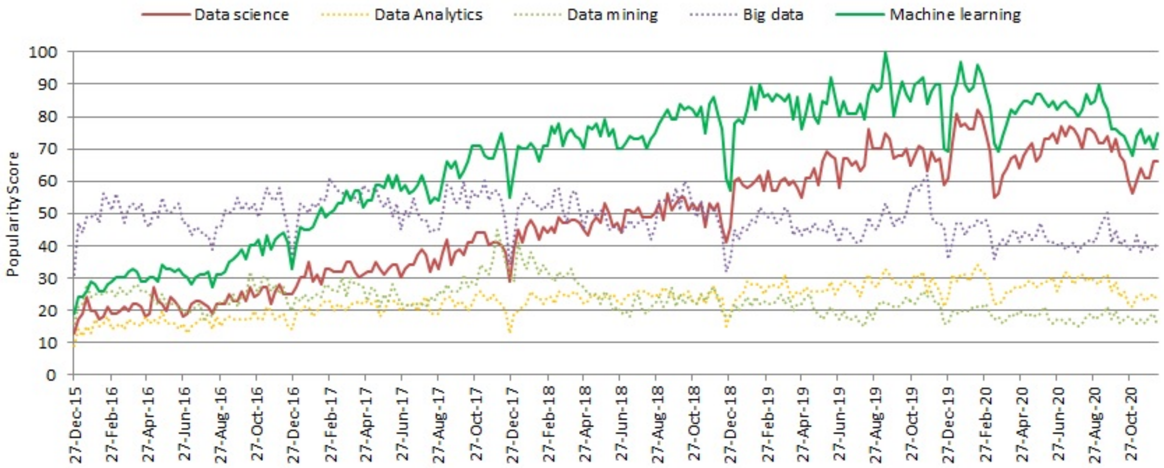

Fig. 1: The worldwide popularity score of data science comparing with relevant topics in a range of 0 (min) to 100 (max) over time where x-axis represents the timestamp information and y-axis represents the corresponding score.

sis of advanced analytics is more on anticipating the use of data to detect patterns to determine what is likely to occur in the future. Basic analytics offer a description of data in general, while *advanced analytics* is a step forward in offering a deeper understanding of data and helping to analyze granular data, which we are interested in. In the field of data science, several types of analytics are popular, such as *Descriptive analytics* which answers the question of what happened; *Diagnostic analytics* which answers the question of why did it happen; *Predictive analytics* which predicts what will happen in the future; and *Prescriptive analytics* which prescribes what action should be taken, discussed briefly in Section 4. Such advanced analytics and decision-making based on *machine learning* techniques [105], a major part of artificial intelligence (AI) [102] can also play a significant role in the Fourth Industrial Revolution (Industry 4.0) due to its learning capability for smart computing as well as automation [121].

Although the area of "data science" is huge, we mainly focus on deriving useful insights through advanced analytics, where the results are used to make smart decisions in various real-world application areas. For this, various *advanced analytics methods* such as machine learning modeling, natural language processing, sentiment analysis, neural network, or deep learning analysis can provide deeper knowledge about data, and thus can be used to develop data-driven intelligent applications. More specifically, regression analysis, classification, clustering analysis, association rules, time-series analysis, sentiment analysis, behavioral patterns, anomaly detection, factor analysis, log analysis, and deep learning which is originated from the artificial neural network, are taken into account in our study. These machine learning-based advanced analytics methods are discussed briefly in Section 4. Thus, it's important to understand the principles of various advanced analytics methods mentioned above and their applicability to apply in various real-world application areas. For instance, in our earlier paper Sarker et al. [114], we have discussed how data science and machine learning modeling can play a significant role in the domain of cybersecurity for making smart decisions and to provide data-driven intelligent security services. In this paper, we broadly take into account the data science application areas and real-world problems in ten potential domains including the area of business data science, health data science, IoT data science, behavioral data science, urban data science, and so on, discussed briefly in Section 5.

Based on the importance of machine learning modeling to extract the useful insights from the data mentioned above and data-driven smart decision-making, in this paper, we present a comprehensive view on *"Data Science"* including various types of advanced analytics methods that can be applied to enhance the intelligence and the capabilities of an application. The key contribution of this study is thus understanding data science modeling, explaining different analytic methods for solution perspective and their applicability in various real-world *data-driven applications* areas mentioned earlier. Overall, the purpose of this paper is, therefore, to provide a basic guide or reference for those *academia and industry* people who want to study, research, and develop automated and intelligent applications or systems based on smart computing and decision making within the area of data science.

The main contributions of this paper are summarized as follows:

- To define the scope of our study towards data-driven smart computing and decision-making in our real-world life. We also make a brief discussion on the concept of data science modeling from business problems to data product and automation, to understand its applicability and provide intelligent services in real-world scenarios.
- To provide a comprehensive view on data science including advanced analytics methods that can be applied to enhance the intelligence and the capabilities of an application.
- To discuss the applicability and significance of machine learning-based analytics methods in various real-world application areas. We also summarize ten potential real-world application areas, from business to personalized applications in our daily life, where advanced analytics with machine learning modeling can be used to achieve the expected outcome.
- To highlight and summarize the challenges and potential research directions within the scope of our study.

The rest of the paper is organized as follows. Section 2 provides the background and related work and defines the scope of our study. Section 3 presents the concepts of data science modeling for building a data-driven application. We briefly discuss and explain different advanced analytics methods and smart computing in Section 4. Various real-world application areas are discussed and summarized in Section 5. In section 6, we highlight and summarize several research issues and potential future directions, and finally, Section 7 concludes this paper.

2 Background and Related Work

In this section, we first discuss various data terms and works related to data science and highlight the scope of our study.

2.1 Data Terms and Definitions

There is a range of key terms in the field, such as data analysis, data mining, data analytics, big data, data science, advanced analytics, machine learning, and deep learning, which are highly related and easily confusing. In the following, we define these terms and differentiate them with the term “Data Science” according to our goal.

The term “Data analysis” refers to the processing of data by conventional (e.g., classic statistical, empirical, or logical) theories, technologies, and tools for extracting useful information and for practical purposes [19]. The term “Data analytics”, on the other hand, refers to the theories, technologies, instruments, and processes that allow for an in-depth understanding and exploration of actionable data insight [19]. Statistical and mathematical analysis of the data is the major concern in this process. “Data mining” is another popular term over the last decade, which has a similar meaning with several other terms such as knowledge mining from data, knowledge extraction, knowledge discovery from data (KDD), data/pattern analysis, data archaeology, and data dredging. According to Han et al. [38], it should have been more appropriately named “knowledge mining from data”. Overall, data mining is defined as the process of discovering interesting patterns and knowledge from large amounts of data [38]. Data sources may include databases, data centers, the Internet or Web, other repositories of data, or data dynamically streamed through the system. “Big data” is another popular term nowadays, which may change the statistical and data analysis approaches as it has the unique features of “massive, high dimensional, heterogeneous, complex, unstructured, incomplete, noisy, and erroneous” [74]. Big data can be generated by mobile devices, social networks, the Internet of Things, multimedia, and many other new applications [129]. Several unique features including Volume, Velocity, Variety, Veracity, Value (5Vs), and complexity are used to understand and describe big data [69].

In terms of analytics, basic analytics provides a summary of data whereas the term “Advanced Analytics” takes a step forward in offering a deeper understanding of data and helps to analyze granular data. Advanced analytics is characterized or defined as autonomous or semi-autonomous data or content analysis using advanced techniques and methods to discover deeper insights, predict or generate recommendations, typically beyond traditional business intelligence or analytics. “Machine learning”, a branch of artificial intelligence (AI), is one of the major techniques used in advanced analytics which can automate analytical model building [112]. This is focused on the premise that systems can learn from data, recognize trends, and make decisions, with minimal human involvement [115] [38]. “Deep Learning” is a subfield of machine learning that discusses algorithms inspired by the human brain’s structure and the function called artificial neural networks [38] [139].

Unlike the above data-related terms, “Data science” is an umbrella term that encompasses advanced data analytics, data mining, machine, and deep learning modeling, and several other related disciplines like statistics, to extract insights or useful knowledge from the datasets and transform them into actionable business strategies. In [19], Cao et al. defined data science from the disciplinary perspective as “data science is a new interdisciplinary field that synthesizes and builds on statistics, informatics, computing, communication, management, and sociology to study data and its environments (including domains and other contextual aspects, such as organizational and social aspects) to transform data to insights and decisions by following a data-to-knowledge-to-wisdom thinking and methodology”. In Section 3, we briefly discuss the data science modeling from a practical perspective starting from business problems to data products that can assist the data scientists to think and work in a particular real-world problem domain within the area of data science and analytics.

2.2 Related Work

In the area, several papers have been reviewed by the researchers based on data science and its significance. For example, the authors in [21] identify the evolving field of data science and its importance in the broader knowledge environment and some issues that differentiate data science and informatics issues from conventional approaches in information sciences. Donoho et al. [28] present 50 years of data science including recent commentary on data science in mass media, and on how/whether data science varies from statistics. The authors formally conceptualize the theory-guided data science (TGDS) model in [53] and present a taxonomy of research themes in TGDS. Cao et al. include a detailed survey and tutorial on the fundamental aspects of data science in [19], which considers the transition from data analysis to data science, the principles of data science, as well as the discipline and competence of data education.

Besides, the authors include a data science analysis in [22], which aims to provide a realistic overview of the use of statistical features and related data science methods in bioimage informatics. The authors in [61] study the key streams of data science algorithm use at central banks and show how their popularity has risen over time. This research contributes to the creation of a research vector on the role of data science in central banking. In [62], the authors provide an overview and tutorial on the data-driven design of intelligent wireless networks. The authors in [87] provide a thorough understanding of computational optimal transport with application to data science. In [97], the authors present data science as theoretical contributions in information systems via text analytics.

Unlike the above recent studies, in this paper, we concentrate on the knowledge of *data science* including advanced analytics methods, machine learning modeling, real-world application domains, and potential research directions within the scope of our study. The advanced analytics methods based on machine learning techniques discussed in this paper can be applied to enhance the capabilities of an application in terms of data-driven intelligent decision making and automation in the final data product or systems.

3 Understanding Data Science Modeling

In this section, we briefly discuss how data science can play a significant role in the real-world business process. For this, we first categorize various types of data and then discuss the major steps of data science modeling starting from business problems to data product and automation.

3.1 Types of Real-World Data

Typically, to build a data-driven real-world system in a particular domain, the availability of data is the key [19] [114] [112]. The data can be in different types such as (i) Structured - that has a well-defined data structure and follows a standard order, examples are names, dates, addresses, credit card numbers, stock information, geolocation, etc.; (ii) Unstructured - has no pre-defined format or organization, examples are sensor data, emails, blog entries, wikis, and word processing documents, PDF files, audio files, videos, images, presentations, web pages, etc.; (iii) Semi-structured - has elements of both the structured and unstructured data containing certain organizational properties, examples are HTML, XML, JSON documents, NoSQL databases, etc.; and (iv) Metadata - that represents data about the data, examples are author, file type, file size, creation date and time, last modification date and time, etc. [105] [38].

In the area of data science, researchers use various widely-used datasets for different purposes. These are, for example, cybersecurity datasets such as NSL-KDD [127], UNSW-NB15 [79], Bot-IoT [59], ISCX’12 [1], CIC-DDoS2019 [2], etc., smartphone datasets such as phone call logs [88] [110], mobile application usages logs [149] [124], SMS Log [29], mobile phone notification logs [77] etc., IoT data [56] [64] [14], health data such as heart disease [99], diabetes mellitus [147] [86], COVID-19 [41] [78], etc., agriculture and e-commerce data [150] [128], and many more in various application domains. In section 5, we discuss ten potential real-world application domains of data science and analytics by taking into account data-driven smart computing and decision

doi:10.20944/preprints202104.0442.v1

making, which can help the data scientists and application developers to explore more in various real-world issues.

Overall, the data used in data-driven applications can be any of the types mentioned above, and they can differ from one application to another in the real world. Data science modeling, which is briefly discussed below, can be used to analyze such data in a specific problem domain and derive insights or useful information from the data to build a data-driven model or data product.

3.2 Steps of Data Science Modeling

Data science is typically an umbrella term that encompasses advanced data analytics, data mining, machine, and deep learning modeling, and several other related disciplines like statistics, to extract insights or useful knowledge from the datasets and transform them into actionable business strategies, mentioned earlier in Section 2. In this section, we briefly discuss how data science can play a significant role in the real-world business process. Figure 2 shows an example of data science modeling starting from real-world data to data-driven product and automation. In the following, we briefly discuss each module of the data science process.

- *Understanding Business Problems:* This involves getting a clear understanding of the problem that is needed to solve, how it impacts the relevant organization or individuals, the ultimate goals for addressing it, and the relevant project plan. Thus to understand and identify the business problems, the data scientists formulate relevant questions while working with the end-users and other stakeholders. For instance, how much/many, which category/group, is the behavior unrealistic/abnormal, which option should be taken, what action, etc. could be relevant questions depending on the nature of the problems. This helps to get a better idea of what business needs and what we should be extracted from data. Such business knowledge can enable organizations to enhance their decision-making process, is known as "Business Intelligence" [65]. Identifying the relevant data sources that can help to answer the formulated questions and what kinds of actions should be taken from the trends that the data shows, is another important task associated with this stage. Once the business problem has been clearly stated, the data scientist can define the analytic approach to solve the problem.
- *Understanding Data:* As we know that data science is largely driven by the availability of data [114]. Thus a sound understanding of the data is needed towards a data-driven model or system. The reason is that real-world data sets are often noisy, missing values, have inconsistencies, or other data issues, which are needed to handle effectively [101]. To gain actionable insights, the appropriate data or the quality of the data must be sourced and cleansed, which is fundamental to any data science engagement. For this, data assessment that evaluates what data is available and how it aligns to the business problem could be the first step in data understanding. Several aspects such as data type/format, the quantity of data whether it is sufficient or not to extract the useful knowledge, data relevance, authorized access to data, feature or attribute importance, combining multiple data sources, important metrics to report the data, etc. are needed to take into account to clearly understand the data for a particular business problem. Overall, the data understanding module involves figuring out what data would be best needed and the best ways to acquire it.
- *Data Exploration and Preparation:* Exploratory data analysis is defined in data science as an approach to analyzing datasets to summarize their key characteristics, often with visual methods [135]. This examines a broad data collection to discover initial trends, attributes, points of interest, etc. in an unstructured manner to construct meaningful summaries of the data. Thus data exploration is typically used to figure out the gist of data and to develop a first step assessment of its quality, quantity, and characteristics. A statistical model can be used or not, but primarily it offers tools for creating hypotheses by generally visualizing and interpreting the data through graphical representation such as a chart, plot, histogram, etc [91] [72]. Before the data is ready for modeling, it's necessary to use data summarization and visualization to audit the quality of the data and provide the information needed to process it. To ensure the quality of the data, the data preparation technique, which is typically the process of cleaning and transforming raw data [107] before processing and analysis is important. It also involves reformatting information, making data corrections, and merging data sets to enrich data. Thus, several aspects such as expected data, data cleaning, formatting or transforming data, dealing with missing values, handling data imbalance and bias issues, data distribution, search for outliers or anomalies in data, ensuring data quality, etc. could be the key considerations in this step.
- *Machine learning Modeling and Evaluation:* Once the data is prepared for building the model, data

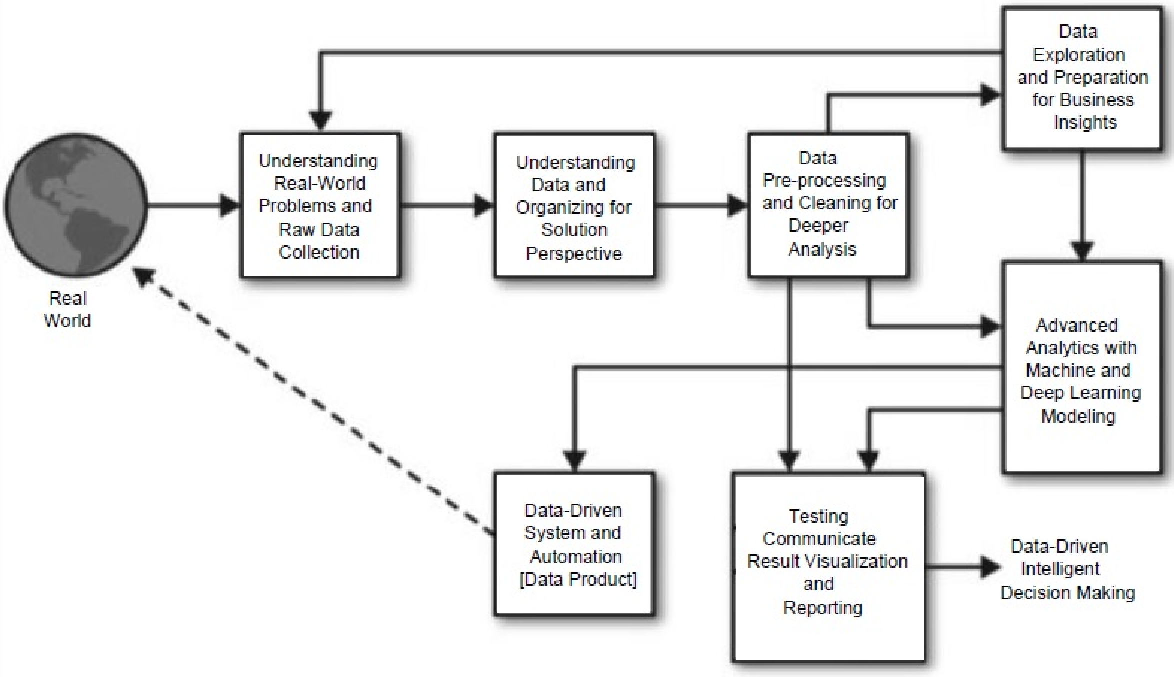
Real World
Understanding Real-World Problems and Raw Data Collection
Understanding Data and Organizing for Solution Perspective
Data Pre-processing and Cleaning for Deeper Analysis
Data Exploration and Preparation for Business Insights
Advanced Analytics with Machine and Deep Learning Modeling
Data-Driven System and Automation [Data Product]
Testing Communicate Result Visualization and Reporting
Data-Driven Intelligent Decision Making

Fig. 2: An example of data science modeling from real-world data to data-driven system and decision making.

scientists design a model or set of models, to address the business problem. Model building is dependent on what type of analytics, e.g., predictive analytics, is needed to solve the particular problem, which is discussed briefly in Section 4. To best fits the data according to the type of analytics, different types of data-driven or machine learning models that have been summarized in our earlier paper Sarker et al. [105] can be built to achieve the goal. Data scientists typically separate training and test subsets of the given dataset usually dividing in the ratio of 80:20 or data considering the most popular $k-$folds data splitting method [38]. This is to observe whether the model performs well or not on the data, to maximize the model performance. Various model validation and assessment metrics, such as error rate, accuracy, true positive, false positive, true negative, false negative, precision, recall, f-score, ROC (receiver operating characteristic curve) analysis, applicability analysis, etc. [115] [38] are used to measure the model performance, which can guide the data scientists to choose or design the learning method or model. Besides, machine learning experts or data scientists can take into account several advanced analytics such as feature engineering, feature selection or extraction methods, algorithm tuning, ensemble methods, modifying existing algorithms, or designing new algorithms, etc. to improve the ultimate data-driven model to solve a particular business problem through smart decision making.

– *Data Product and Automation:* A data product is typically the output of any data science activity [19]. A data product, in general terms, is a data deliverable, or data-enabled or guide, which can be a discovery, prediction, service, suggestion, insight into decision-making, thought, model, paradigm, tool, application, or system that process data and generate results. Businesses can use the results of such data analysis to obtain useful information like churn (a measure of how many customers stop using a product) prediction and customer segmentation, and use these results to make smarter business decisions and automation. Thus to make better decisions in various business problems, various machine learning pipelines and data products can be developed. To highlight this, we summarize several potential real-world data science application areas in Section 5, where various data products can play a significant role in relevant business problems to make them smart and automate.

Overall, we can conclude that data science modeling can be used to help drive changes and improvements in business practices. The interesting part of the data science process indicates having a deeper understanding of the business problem to solve. Without that, it would be much harder to gather the right data and extract the most useful information from the data for making decisions to solve the problem. In terms of role, "Data Scientists" typically interpret and manage data to uncover the answers to major questions that help organizations to make objective decisions and solve complex problems. In a summary, a data scientist proactively gathers and analyzes information from multiple sources to better understand how the business performs, and develops machine learning or data-driven tools focused on advanced analytics, which can make today's computing process smarter and intelligent, discussed briefly in the following section.

4 Advanced Analytics Methods and Smart Computing

As mentioned earlier in Section 2, basic analytics provides a summary of data whereas advanced analytics takes a step forward in offering a deeper understanding of data and helps in granular data analysis. For instance, the predictive capabilities of advanced analytics can be used to forecast trends, events, and behaviors. Thus, "advanced analytics" can be defined as the autonomous or semi-autonomous analysis of data or content using advanced techniques and methods to discover deeper insights, make predictions, or produce recommendations, where machine learning-based analytical modeling is considered as the key technologies in the area. In the following section, we first summarize various types of analytics and outcome that are needed to solve the associated business problems, and then we briefly discuss machine learning-based analytical modeling.

4.1 Types of Analytics and Outcome

In the real-world business process, several key questions such as "what happened?", "Why did it happen?", "What will happen in the future?", "What action should be taken?" are common and important. Based on these questions, in this paper, we categorize and highlight the analytics into four types such as descriptive, diagnostic, predictive, and prescriptive, which are discussed below.

– *Descriptive Analytics:* It is the interpretation of historical data to better understand the changes that have occurred in a business. Thus descriptive analytics answers the question, "what happened in the past?" by summarizing past data such as statistics on sales and operations or marketing strategies, use of social media, and engagement with Twitter, Linkedin or Facebook, etc. For instance, using descriptive analytics through analyzing trends, patterns, and anomalies, etc., customers' historical shopping data can be used to predict the probability of a customer purchasing a product. Thus, descriptive analytics can play a significant role to provide an accurate picture of what has occurred in a business and how it relates to previous times utilizing a broad range of relevant business data. As a result, managers and decision-makers can pinpoint areas of strength and weakness in their business, and eventually can take more effective management strategies and business decisions.
– *Diagnostic Analytics:* It is a form of advanced analytics that examines data or content to answer the question, "why did it happen?" The goal of diagnostic analytics is to help to find the root cause of the problem. For example, the human resource management department of a business organization may use these diagnostic analytics to find the best applicant for a position, select them, and compare them to other similar positions to see how well they perform. In a healthcare example, it might help to figure out whether the patients' symptoms such as high fever, dry cough, headache, fatigue, etc. are all caused by the same infectious agent. Overall, diagnostic analytics enables one to extract value from the data by posing the right questions and conducting in-depth investigations into the answers. It is characterized by techniques such as drill-down, data discovery, data mining, and correlations.
– *Predictive Analytics:* Predictive analytics is an important analytical technique used by many organizations for various purposes such as to assess business risks, anticipate potential market patterns, and decide when maintenance is needed, to enhance their business. It is a form of advanced analytics that examines data or content to answer the question, "what will happen in the future?" Thus, the primary goal of predictive analytics is to identify and typically answer this question with a high degree of probability. Data scientists can use historical data as a source to extract insights for building predictive models using various regression analyses and machine learning techniques, which can be used in various application domains for a better outcome. Companies, for example, can use predictive analytics to minimize costs by better anticipating future demand and changing output and inventory, banks and other financial institutions to reduce fraud and risks by predicting suspicious activity, medical specialists to make effective decisions through predicting patients who are at risk of diseases, retailers to increase sales and customer satisfaction through understanding and predicting customer preferences,

Table 1: Various types of analytical methods with examples.

Analytical Methods	Data-Driven Model Building	Examples
Descriptive Analytics	Answer the question, "what happened in the past"?	Summarising past events, e.g., sales, business data, social media usage, reporting general trends, etc.
Diagnostic Analytics	Answer the question, "why did it happen?"	Identify anomalies and determine casual relationships, to find out business loss, identifying the influence of medications, etc.
Predictive Analytics	Answer the question, "what will happen in the future?"	Predicting customer preferences, recommending products, identifying possible security breaches, predicting staff and resource needs, etc.
Prescriptive Analytics	Answer the question, "what action should be taken?"	Improving business management, maintenance, improving patient care and healthcare administration, determining optimal marketing strategies, etc.

manufacturers to optimize production capacity through predicting maintenance requirements, and many more. Thus predictive analytics can be considered as the core analytical method within the area of data science.

- *Prescriptive Analytics:* Prescriptive analytics focuses on recommending the best way forward with actionable information to maximize overall returns and profitability, which typically answer the question, "what action should be taken?" In business analytics, prescriptive analytics is considered the final step. For its models, prescriptive analytics collects data from several descriptive and predictive sources and applies it to the decision-making process. Thus, we can say that it is related to both descriptive analytics and predictive analytics, but it emphasizes actionable insights instead of data monitoring. In other words, it can be considered as the opposite of descriptive analytics, which examines decisions and outcomes after the fact. By integrating big data, machine learning, and business rules, prescriptive analytics helps organizations to make more informed decisions to produce results that drive the most successful business decisions.

In summary, to clarify what happened and why it happened, both descriptive analytics and diagnostic analytics look at the past. Historical data is used by predictive analytics and prescriptive analytics to forecast what will happen in the future and what steps should be taken to impact those effects. In Table 1, we have summarized these analytics methods with examples. Forward-thinking organizations in the real world can jointly use these analytical methods to make smart decisions that help drive changes in business processes and improvements. In the following, we discuss how machine learning techniques can play a big role in these analytical methods through their learning capabilities from the data.

4.2 Machine Learning based Analytical Modeling

In this section, we briefly discuss various advanced analytics methods based on machine learning modeling, which can make the computing process smart through intelligent decision-making in a business process. Fig. 3 shows a general structure of a machine learning-based predictive modeling considering both the training and testing phase. In the following, we discuss a wide range of methods such as regression and classification analysis, association rule analysis, time-series analysis, behavioral analysis, log analysis, and so on within the scope of our study.

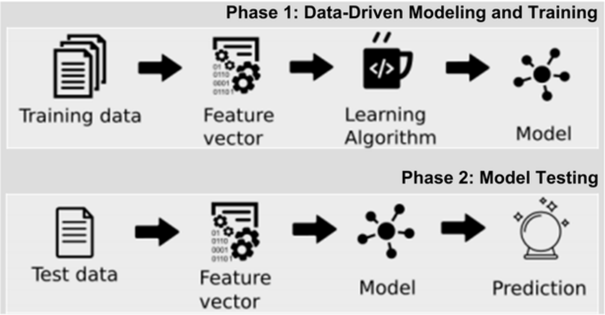

Fig. 3: A general structure of a machine learning based predictive model considering both the training and testing phase.

4.2.1 Regression Analysis

In data science, one of the most common statistical approaches used for predictive modeling and data mining tasks is regression techniques [38]. Regression analysis is a form of supervised machine learning that examines the relationship between a dependent variable (target) and independent variables (predictor) in order to predict continuous-valued output [117] [105]. The following equations Equ. 1, Equ. 2, and Equ. 3 [85] [105] represent the simple, multiple or multivariate, and polynomial regressions respectively, where x represents independent variable and y is the predicted/target output mentioned above.

$$y = a + bx + e \tag{1}$$

$$y = a + b_1x_1 + b_2x_2 + ... + b_nx_n + e \tag{2}$$

$$y = a + b_1x + b_2x^2 + b_3x^3 + ... + b_nx^n + e \tag{3}$$

Regression analysis is typically conducted for one of two purposes: to predict the value of the dependent variable in the case of individuals for whom some knowledge relating to the explanatory variables is available, or to estimate the effect of some explanatory variable on the dependent variable, i.e., finding the relationship of causal influence between the variables. Linear regression cannot be used to fit non-linear data and may cause an underfitting problem. In that case, polynomial regression performs better, however, increases the model complexity. The regularization techniques such as Ridge, Lasso, Elastic-Net, etc. [85] [105] can be used to optimize the linear regression model. Besides, Support Vector Regression, Decision Tree Regression, Random Forest Regression techniques [85] [105] can be used for building effective regression models depending on the problem type, e.g., non-linear tasks. Financial forecasting or prediction, cost estimation, trend analysis, marketing, time-series estimation, drug response modeling, etc. are some examples where the regression models can be used to solve real-world problems in the domain of data science and analytics.

4.2.2 Classification Analysis

Classification is one of the most widely used and best-known data science processes. This is a form of supervised machine learning approach that also refers to a predictive modeling problem in which a class label is predicted for a given example [38]. Spam identification, such as 'spam' and 'not spam' in email service providers, can be an example of a classification problem. There are several forms of classification analysis available in the area such as binary classification - which refers to the prediction of one of two classes; Multi-class classification - which involves the prediction of one of more than two classes; Multi-label classification - a generalization of multiclass classification in which the problem's classes are organized hierarchically [105].

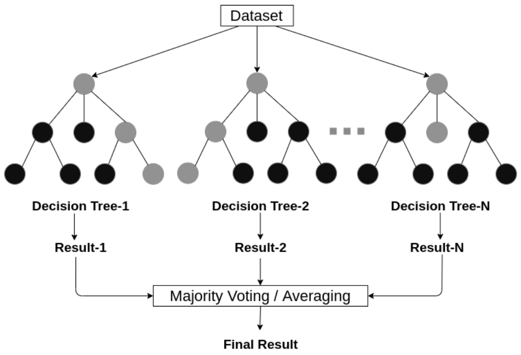

Fig. 4: An example of a random forest structure considering multiple decision trees.

Several popular classification techniques, such as k-nearest neighbors [8], support vector machines [55], navies Bayes [49], adaptive boosting [33], extreme gradient boosting [85], logistic regression [66], decision trees ID3 [92], C4.5 [93], and random forests [16] exist to solve classification problems. The tree-based classification technique, e.g., random forest considering multiple decision trees, performs better than others to solve real-world problems in many cases as due to its capability of producing logic rules [103] [115]. Figure 4 shows an example of a random forest structure considering multiple decision trees. In addition, BehavDT recently proposed by Sarker et al. [109], and IntrudTree [106] can be used for building effective classification or prediction models in the relevant tasks within the domain of data science and analytics.

4.2.3 Cluster Analysis

Clustering is a form of unsupervised machine learning technique and is well-known in many data science application areas for statistical data analysis [38]. Usually, clustering techniques search for the structures inside a dataset and, if the classification is not previously identified, classify homogeneous groups of cases. This means that data points are identical to each other within a cluster, and different from data points in another cluster. Overall, the purpose of cluster analysis is to sort various data points into groups (or clusters) that are homogeneous internally and heterogeneous externally [105]. To gain insight into how data is distributed in a given dataset or as a preprocessing phase for other algorithms, clustering is often used. Data clustering, for example, assists with customer shopping behavior, sales campaigns, and retention of consumers for retail businesses, anomaly detection, etc.

Many clustering algorithms with the ability to group data have been proposed in machine learning and data science literature [141] [98] [138]. In our earlier paper Sarker et al. [105], we have summarized this based on several perspectives, such as partitioning methods, density-based methods, hierarchical-based methods, model-based methods, etc. In the literature, the popular K-means [75], K-Mediods [84], CLARA [54] etc. are known as partitioning methods; DBSCAN [31], OPTICS [11] etc. are known as density-based methods; single linkage [122], complete linkage [123], etc. are known as hierarchical methods. In addition, grid-based clustering methods, such as STING [134], CLIQUE [5], etc.; model-based clustering such as neural network learning [141], GMM [94], SOM [20] [104], etc.; constrained-based methods such as COP K-means [131], CMWK-Means [26], etc. are used in the area. Recently, Sarker et al. [111] proposed a hierarchical clustering method, BOTS [111] based on bottom-up agglomerative technique for capturing user's similar behavioral characteristics over time. The key benefit of agglomerative hierarchical clustering is that the tree-structure hierarchy created by agglomerative clustering is more informative than an unstructured set of flat clusters, which can assist in better decision-making in relevant application areas in data science.

4.2.4 Association Rule Analysis

Association rule learning is known as a rule-based machine learning system, an unsupervised learning method is typically used to establish a relationship among variables. This is a descriptive technique often used to analyze large datasets for discovering interesting relationships or patterns. The association learning technique's main strength is its comprehensiveness, as it produces all associations that meet user-specified constraints including minimum support and confidence value [138].

Association rules allow a data scientist to identify trends, associations, and co-occurrences between data sets inside large data collections. In a supermarket, for example, associations infer knowledge about the buying behavior of consumers for different items, which helps to change the marketing and sales plan. In healthcare, to better diagnose patients, physicians may use association guidelines. Doctors can assess the conditional likelihood of a given illness by comparing symptom associations in the data from previous cases by using association rules and machine learning-based data analysis. Similarly, association rules are useful for consumer behavior analysis and prediction, customer market analysis, bioinformatics, weblog mining, recommendation systems, etc.

Several types of association rules have been proposed in the area, such as frequent pattern based [7] [47], [73], logic-based [32], tree-based [39], fuzzy-rules [126], belief rule [148] etc. The rule learning techniques such as AIS [6], Apriori [7], Apriori-TID and Apriori-Hybrid [7], FP-Tree [39], Eclat [144], RARM [25] exist to solve the relevant business problems. Apriori [7] is the most commonly used algorithm for discovering association rules from a given dataset among the association rule learning techniques [145]. The recent association rule-learning technique ABC-RuleMiner proposed in our earlier paper by Sarker et al. [113] could give significant results in terms of generating non-redundant rules that can be used for smart decision making according to human preferences, within the area of data science applications.

4.2.5 Time-series Analysis and Forecasting

A time series is typically a series of data points indexed in time order particularly, by date, or timestamp [111]. Depending on the frequency, the time-series can be different types such as annually, e.g., annual budget, quarterly, e.g., expenditure, monthly, e.g., air traffic, weekly, e.g., sales quantity, daily, e.g., weather, hourly, e.g., stock price, minute-wise, e.g., inbound calls in a call center, and even second-wise, e.g., web traffic, and so on in relevant domains.

A mathematical method dealing with such time-series data, or the procedure of fitting a time series to a proper model is termed time-series analysis. Many different time series forecasting algorithms and analysis methods can be applied to extract the relevant information. For instance, to do time-series forecasting for future patterns, the Autoregressive (AR) model [130] learns the behavioral trends or patterns of past data. Moving average (MA) [40] is another simple and common form of smoothing used in time series analysis and forecasting that uses past forecasted errors in a regression-like model to elaborate an averaged trend across the data. The Autoregressive Moving Average (ARMA) [15] [120] combines these two approaches, where Autoregressive extracts the momentum and pattern of the trend and Moving Average capture the noise effects. The most popular and frequently used time-series model is the Autoregressive Integrated Moving Average (ARIMA) model [15] [120]. ARIMA model, a generalization of an ARMA model, is more flexible than other statistical models such as exponential smoothing or simple linear regression. In terms of data, the ARMA model can only be used for stationary time-series data, while the ARIMA model includes the case of non-stationarity as well. Similarly, Seasonal Autoregressive Integrated Moving Average (SARIMA), Autoregressive Fractionally Integrated Moving Average (ARFIMA), Autoregressive Moving Average model with exogenous inputs model (ARMAX model) are also used in time-series models [120].

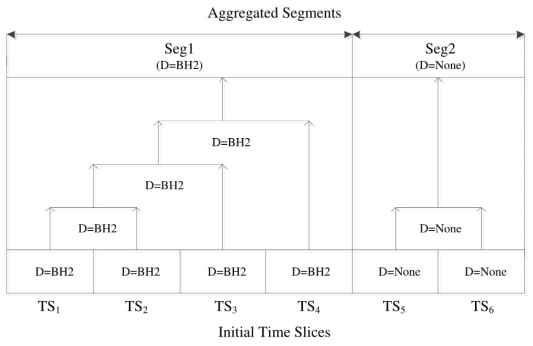

Fig. 5: An example of producing aggregate time segments from initial time slices based on similar behavioral characteristics.

In addition to the stochastic methods for time-series modeling and forecasting, machine and deep learning-based approach can be used for effective time-series analysis and forecasting. For instance, in our earlier paper, Sarker et al. [111] present a bottom-up clustering-based time-series analysis to capture the mobile usage behavioral patterns of the users. Figure 5 shows an example of producing aggregate time segments from initial time slices based on similar behavioral characteristics that are used in our bottom-up clustering approach mentioned above [111]. The authors in [118], used a Long Short-Term Memory (LSTM) model, a kind of Recurrent Neural Network (RNN) deep learning model, in forecasting time-series that outperform traditional approaches such as the ARIMA model. Time-series analysis is commonly used these days in various fields such as financial, manufacturing, business, social media, event data (e.g., clickstreams and system events), IoT and smartphone data, and generally in any applied science and engineering temporal measurement domain. Thus, it covers a wide range of application areas in data science.

4.2.6 Opinion Mining and Sentiment Analysis

Sentiment analysis or opinion mining is the computational study of the opinions, thoughts, emotions, assessments, and attitudes of people towards entities such as products, services, organizations, individuals, issues, events, topics, and their attributes [71]. There are three kinds of sentiments: positive, negative, and neutral, along with more extreme feelings such as angry, happy and sad, or interested or not interested, etc. More refined sentiments to evaluate the feelings of individuals in various situations can also be found according to the problem domain.

Although the task of opinion mining and sentiment analysis is very challenging from a technical point of view, it's very useful in real-world practice. For instance, a business always aims to obtain an opinion from the public or customers about its products and services to refine the business policy as well as a better business decision. It can thus benefit a business to understand the social opinion of their brand, product, or service. Besides, potential customers want to know what consumers believe they have when they use a service or purchase a product. Document-level, sentence level, aspect level, and concept level, are the possible levels of opinion mining in the area [45].

Several popular techniques such as lexicon-based including dictionary-based and corpus-based methods, machine learning including supervised and unsupervised learning, deep learning, and hybrid methods are used in sentiment analysis-related tasks [70]. To systematically define, extract, measure, and analyze affective states and subjective knowledge, it incorporates the use of statistics, natural language processing (NLP), machine learning as well as deep learning methods. Sentiment analysis is widely used in many applications, such as reviews and survey data, web and social media, and healthcare content, ranging from marketing and customer support to clinical practice. Thus sentiment analysis has a big influence in many data science applications, where public sentiment is involved in various real-world issues.

4.2.7 Behavioral Data and Cohort Analysis

Behavioral analytics is a recent trend that typically reveals new insights into e-commerce sites, online gaming, mobile and smartphone applications, IoT user behavior, and many more [112]. The behavioral analysis aims to understand how and why the consumers or users behave, allowing accurate predictions of how they are likely to behave in the future. For instance, it allows advertisers to make the best offers with the right client segments at the right time. Behavioral analytics, including traffic data such as navigation paths, clicks, social media interactions, purchase decisions, and marketing responsiveness, use the large quantities of raw user event information gathered during sessions in which people use apps, games, or websites. In our earlier paper Sarker et al. [111] [101] [113] we have discussed how to

doi:10.20944/preprints202104.0442.v1

extract users phone usage behavioral patterns utilizing real-life phone log data for various purposes.

In the real-world scenario, behavioral analytics is often used in e-commerce, social media, call centers, billing systems, IoT systems, political campaigns, and other applications, to find opportunities for optimization to achieve particular outcomes. Cohort analysis is a branch of behavioral analytics that involves studying groups of people over time to see how their behavior changes. For instance, it takes data from a given data set (e.g., an e-commerce website, web application, or online game) and separates it into related groups for analysis. Various machine learning techniques such as behavioral data clustering [111], behavioral decision tree classification [109], behavioral association rules [113], etc. can be used in the area according to the goal. Besides, the concept of RecencyMiner, proposed in our earlier paper Sarker et al. [108] that takes into account recent behavioral patterns could be effective while analyzing behavioral data as it may not be static in the real-world changes over time.

4.2.8 Anomaly Detection or Outlier Analysis

Anomaly detection, also known as Outlier analysis is a data mining step that detects data points, events, and/or findings that deviate from the regularities or normal behavior of a dataset. Anomalies are usually referred to as outliers, abnormalities, novelties, noise, inconsistency, irregularities, and exceptions [63] [114]. Techniques of anomaly detection may discover new situations or cases as deviant based on historical data through analyzing the data patterns. For instance, identifying fraud or irregular transactions in finance is an example of anomaly detection.

It is often used in preprocessing tasks for the deletion of anomalous or inconsistency in the real-world data collected from various data sources including user logs, devices, networks, and servers. For anomaly detection, several machine learning techniques can be used, such as k-nearest neighbors, isolation forests, cluster analysis, etc [105]. The exclusion of anomalous data from the dataset also results in a statistically significant improvement in accuracy during supervised learning [101]. However, extracting appropriate features, identifying normal behaviors, managing imbalanced data distribution, addressing variations in abnormal behavior or irregularities, the sparse occurrence of abnormal events, environmental variations, etc. could be challenging in the process of anomaly detection. Detection of anomalies can be applicable in a variety of domains such as cybersecurity analytics, intrusion detections, fraud detection, fault detection, health analytics, identifying irregularities, detecting ecosystem disturbances, and many more. This anomaly detection can be considered a significant task for building effective systems with higher accuracy within the area of data science.

4.2.9 Factor Analysis

Factor analysis is a collection of techniques for describing the relationships or correlations between variables in terms of more fundamental entities known as factors [24]. It's usually used to organize variables into a small number of clusters based on their common variance, where mathematical or statistical procedures are used. The goals of factor analysis are to determine the number of fundamental influences underlying a set of variables, calculate the degree to which each variable is associated with the factors, and learn more about the existence of the factors by examining which factors contribute to output on which variables. The broad purpose of factor analysis is to summarize data so that relationships and patterns can be easily interpreted and understood [143].

Exploratory Factor Analysis (EFA) and Confirmatory Factor Analysis (CFA) are the two most popular factor analysis techniques. EFA seeks to discover complex trends by analyzing the dataset and testing predictions, while CFA tries to validate hypotheses and uses path analysis diagrams to represent variables and factors [143]. Factor analysis is one of the algorithms for unsupervised machine learning that is used for minimizing dimensionality. The most common methods for factor analytics are Principal Components Analysis (PCA), Principal Axis Factoring (PAF), and Maximum Likelihood (ML) [48]. Methods of correlation analysis such as Pearson correlation, canonical correlation, etc. may also be useful in the field as they can quantify the statistical relationship between two continuous variables, or association. Factor analysis is commonly used in finance, marketing, advertising, product management, psychology, and operations research, and thus can be considered as another significant analytical method within the area of data science.

4.2.10 Log Analysis

Logs are commonly used in system management as logs are often the only data available that record detailed system runtime activities or behaviors in production [44]. Log analysis is thus can be considered as the method of analyzing, interpreting, and capable of understanding computer-generated records or messages, also known as logs. This can be device log, server log, system log,

doi:10.20944/preprints202104.0442.v1

network log, event log, audit trail, audit record, etc. The process of creating such records is called data logging.

Logs are generated by a wide variety of programmable technologies, including networking devices, operating systems, software, and more. Phone call logs [88] [110], SMS Logs [29], mobile apps usages logs [149] [124], notification logs [77], game Logs [82], context logs [18] [149], web logs [37], smartphone life logs [95], etc. are some examples of log data for smartphone devices. The main characteristics of these log data is that it contains users' actual behavioral activities with their devices. Similar other log data can be search logs [133] [50], application logs [27], server logs [34], network logs [57], event logs [83], network and security logs [142] etc.

Several techniques such as classification and tagging, correlation analysis, pattern recognition methods, anomaly detection methods, machine learning modeling, etc. [105] can be used for effective log analysis. Log analysis can assist in compliance with security policies and industry regulations, as well as provide a better user experience by encouraging the troubleshooting of technical problems and identifying areas where efficiency can be improved. For instance, web servers use log files to record data about website visitors. Windows event log analysis can help an investigator draw a timeline based on the logging information and the discovered artifacts. Overall, advanced analytics methods by taking into account machine learning modeling can play a significant role to extract insightful patterns from these log data, which can be used for building automated and smart applications, and thus can be considered as a key working area in data science.

4.2.11 Neural Networks and Deep Learning analysis

Deep learning is a form of machine learning that uses artificial neural networks to create a computational architecture that learns from data by combining multiple processing layers, such as the input, hidden, and output layers [38]. The key benefit of deep learning over conventional machine learning methods is that it performs better in a variety of situations, particularly when learning from large datasets [114] [140].

The most common deep learning algorithms are: Multi-layer perceptron (MLP) [85], Convolutional Neural Network (CNN or ConvNet) [67], Long Short Term Memory Recurrent Neural Network (LSTM-RNN) [35]. Figure 6 shows a structure of an artificial neural network modeling with multiple processing layers. The Backpropagation technique [38] is used to adjust the weight values internally while building the model. Convolutional neural networks (CNNs) [67] improve on the design of traditional artificial neural networks (ANNs), which include convolutional layers, pooling layers, and fully connected layers. It is commonly used in a variety of fields, including natural language processing, speech recognition, image processing, and other autocorrelated data since it takes advantage of the two-dimensional (2D) structure of the input data. AlexNet [60], Xception [23], Inception [125], Visual Geometry Group (VGG) [42], ResNet [43], etc., and other advanced deep learning models based on CNN are also used in the field.

In addition to CNN, recurrent neural network (RNN) architecture is another popular method used in deep learning. Long short-term memory (LSTM) is a popular type of recurrent neural network architecture used broadly in the area of deep learning. Unlike traditional feed-forward neural networks, LSTM has feedback connections. Thus, LSTM networks are well-suited for analyzing and learning sequential data, such as classifying, sorting, and predicting data based on time-series data. Therefore, when the data is in a sequential format, such as time, sentence, etc., LSTM can be used, and it is widely used in the areas of time-series analysis, natural language processing, speech recognition, and so on.

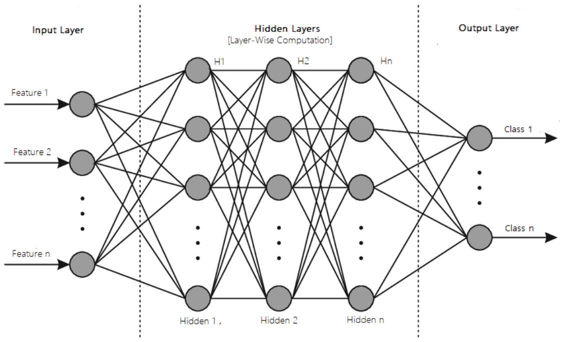

Fig. 6: A structure of an artificial neural network modeling with multiple processing layers.

In addition to the most popular deep learning methods mentioned above, several other deep learning approaches [104] exist in the field for various purposes. The self-organizing map (SOM) [58], for example, uses unsupervised learning to represent high-dimensional data as a 2D grid map, reducing dimensionality. Another learning technique that is commonly used for dimensionality reduction and feature extraction in unsupervised learning tasks is the autoencoder (AE) [13]. Restricted Boltzmann machines (RBM) can be used for dimensionality reduction, classification, regression, collaborative filtering, feature learning, and topic modeling, according to [46]. A deep belief network (DBN) is usually made up of a backpropagation neural network and unsupervised networks like restricted Boltzmann machines (RBMs) or autoencoders (BPNN) [136]. A

doi:10.20944/preprints202104.0442.v1

generative adversarial network (GAN) [36] is a deep learning network that can produce data with characteristics that are similar to the input data. Transfer learning is common worldwide presently because it can train deep neural networks with a small amount of data, which is usually the re-use of a pre-trained model on a new problem [137]. The key benefit of deep learning over conventional machine learning methods is that it can perform better when learning from large-scale datasets [105] [140]. In our previous article Sarker et al. [104], we have summarized a brief discussion of various artificial neural networks (ANN) and deep learning (DL) models mentioned above, which can be used in a variety of data science and analytics tasks.

5 Real-World Application Domains

Almost every industry or organization is impacted by data, and thus "Data Science" including advanced analytics with machine learning modeling can be used in business, marketing, finance, IoT systems, cybersecurity, urban management, health care, government policies, and every possible industry where data gets generated. In the following, we discuss several most popular application areas based on data science.

- *Ecommerce or Business Data Science:* This is the study of business data to obtain insights about a business that can typically lead to making improved business decisions and taking high-quality actions. Business data is typically the information that is used to plan and operate an organization [90]. To understand and analyze the actual phenomena with business data for making better business decisions, various advanced analytics methods, algorithms, machine learning techniques, and systems can be used, which we can define as business data science. It can explore the historical business data, compare it to the competition, evaluate the demand, and make decisions on when and where the product or service will sell the best [105]. This can also assist businesses in better understanding their customers, evaluating their advertising campaigns, personalizing marketing, developing a content strategy and products. Many online retailers, such as Amazon [76], can improve inventory management, avoid out-of-stock situations, and optimize logistics and warehousing by using predictive modeling based on machine learning techniques. Overall, business data science technologies can help businesses produce the best products at the right time and create new products or services to meet their customers' needs, making the business process smarter and more intelligent.
- *Manufacturing or Industrial Data Science:* To compete in global production capability, quality, and cost, manufacturing industries have gone through many industrial revolutions [17]. The latest fourth industrial revolution, also known as Industry 4.0, is the emerging trend of automation and data exchange in manufacturing technology. Thus industrial data science, which is the study of industrial data to obtain insights that can typically lead to optimizing industrial applications, can play a vital role in such revolution. Manufacturing industries generate a large amount of data from various sources such as sensors, devices, networks, systems, and applications [9] [68]. The main categories of industrial data include large-scale data devices, life-cycle production data, enterprise operation data, manufacturing value chain sources, and collaboration data from external sources [132]. The data needs to be processed, analyzed, and secured to help improve the system's efficiency, safety, and scalability. Data science modeling thus can be used to maximize production, reduce costs and raise profits in manufacturing industries.
- *Medical or Health Data Science:* Healthcare is one of the most notable fields where data science is making major improvements. Health data science involves the extrapolation of actionable insights from sets of patient data, typically collected from electronic health records (EHRs). To help organizations, improve the quality of treatment, lower the cost of care, and improve the patient experience, data can be obtained from several sources, e.g., the electronic health record, billing claims, cost estimates, and patient satisfaction surveys, etc., to analyze. In reality, healthcare analytics using machine learning modeling can minimize medical costs, predict infectious outbreaks, prevent preventable diseases, and generally improve the quality of life [119] [81]. Across the global population, the average human lifespan is growing, presenting new challenges to today's methods of delivery of care. Thus health data science modeling can play a role in analyzing current and historical data to predict trends, improve services, and even better monitor the spread of diseases. Eventually, it may lead to new approaches to improve patient care, clinical expertise, diagnosis, and management.
- *IoT Data Science:* Internet of Things (IoT) [12] is a revolutionary technical field that turns every electronic system into a smarter one and is therefore considered to be the big frontier that can enhance almost all activities in our lives. Machine learning has become a key technology for IoT applications

doi:10.20944/preprints202104.0442.v1

because it uses expertise to identify patterns and generate models that help predict future behavior and events [112]. One of the IoT's main fields of application is a smart city, which uses technology to improve city services and citizens' living experiences. For example, using the relevant data, data science methods can be used for traffic prediction in smart cities, to estimate the total usage of energy of the citizens for a particular period. Deep learning-based models in data science can be built based on a large scale of IoT datasets [10] [104]. Overall, data science and analytics approaches can aid modeling in a variety of IoT and smart city services, including smart governance, smart homes, education, connectivity, transportation, business, agriculture, health care, and industry, and many others.

- *Cybersecurity Data Science:* Cybersecurity, or the practice of defending networks, systems, hardware, and data from digital attacks, is one of the most important fields of Industry 4.0 [121] [114]. Data science techniques, particularly machine learning, have become a crucial cybersecurity technology that continually learns to identify trends by analyzing data, better detecting malware in encrypted traffic, finding insider threats, predicting where bad neighborhoods are online, keeping people safe while surfing, or protecting information in the cloud by uncovering suspicious user activity [114]. For instance, machine learning and deep learning-based security modeling can be used to effectively detect various types of cyberattacks or anomalies [106] [103]. To generate security policy rules, association rule learning can play a significant role to build rule-based systems [102]. Deep learning-based security models can perform better when utilizing the large scale of security datasets [140]. Thus data science modeling can enable professionals in cybersecurity to be more proactive in preventing threats and reacting in real-time to active attacks, through extracting actionable insights from the security datasets.
- *Behavioral Data Science:* Behavioral data is information produced as a result of activities, most commonly commercial behavior, performed on a variety of Internet-connected devices, such as a PC, tablet, or smartphones [112]. Websites, mobile applications, marketing automation systems, call centers, help desks, and billing systems, etc. are all common sources of behavioral data. Behavioral data is much more than just data, which is not static data [108]. Advanced analytics of these data including machine learning modeling can facilitate in several areas such as predicting future sales trends and product recommendations in e-commerce and retail; predicting usage trends, load, and user preferences in future releases in online gaming; determining how users use an application to predict future usage and preferences in application development; breaking users down into similar groups to gain a more focused understanding of their behavior in cohort analysis; detecting compromised credentials and insider threats by locating anomalous behavior, or making suggestions, etc. Overall, behavioral data science modeling typically enables to make the right offers to the right consumers at the right time on various common platforms such as e-commerce platforms, online games, web and mobile applications, and IoT.
- *Mobile Data Science:* Today's smart mobile phones are considered as "next-generation, multi-functional cell phones that facilitate data processing, as well as enhanced wireless connectivity" [146]. In our earlier paper [112], we have shown that users' interest in "Mobile Phones" is more and more than other platforms like "Desktop Computer", "Laptop Computer" or "Tablet Computer" in recent years. People use smartphones for a variety of activities, including e-mailing, instant messaging, online shopping, Internet surfing, entertainment, social media such as Facebook, Linkedin, and Twitter, and various IoT services such as smart cities, health, and transportation services, and many others. Intelligent apps are based on the extracted insight from the relevant datasets depending on apps characteristics, such as action-oriented, adaptive in nature, suggestive and decision-oriented, data-driven, context-awareness, and cross-platform operation [112]. As a result, mobile data science, which involves gathering a large amount of mobile data from various sources and analyzing it using machine learning techniques to discover useful insights or data-driven trends, can play an important role in the development of intelligent smartphone applications.
- *Multimedia Data Science:* Over the last few years, a big data revolution in multimedia management systems has resulted from the rapid and widespread use of multimedia data, such as image, audio, video, and text, as well as the ease of access and availability of multimedia sources. Currently, multimedia sharing websites, such as Yahoo Flickr, iCloud, and YouTube, and social networks such as Facebook, Instagram, and Twitter, are considered as valuable sources of multimedia big data [89]. People, particularly younger generations, spend a lot of time on the Internet and social networks to connect with others, exchange information, and create multimedia data, thanks to the advent of new

technology and the advanced capabilities of smartphones and tablets. Multimedia analytics deals with the problem of effectively and efficiently manipulating, handling, mining, interpreting, and visualizing various forms of data to solve real-world problems. Text analysis, image or video processing, computer vision, audio or speech processing, and database management are among the solutions available for a range of applications including healthcare, education, entertainment, and mobile devices.

- *Smart Cities or Urban Data Science:* Today, more than half of the world's population live in urban areas or cities [80] and considered as drivers or hubs of economic growth, wealth creation, well-being, and social activity [116] [96]. In addition to cities, "Urban area" can refer to the surrounding areas such as towns, conurbations, or suburbs. Thus, a large amount of data documenting daily events, perceptions, thoughts, and emotions of citizens or people are recorded, that are loosely categorized into personal data, e.g., household, education, employment, health, immigration, crime, etc., proprietary data, e.g., banking, retail, online platforms data, etc., government data, e.g., citywide crime statistics, or government institutions, etc., Open and public data, e.g., data.gov, ordnance survey, and organic and crowdsourced data, e.g., user-generated web data, social media, Wikipedia, etc. [30]. The field of Urban Data Science typically focuses on providing more effective solutions from a data-driven perspective, through extracting knowledge and actionable insights from such urban data. Advanced analytics of these data including machine learning modeling can facilitate the efficient management of urban areas including real-time management, e.g., traffic flow management, evidence-based planning decisions which pertain to the longer-term strategic role of forecasting for urban planning, e.g., crime prevention, public safety, and security, or framing the future, e.g., political decision-making [30]. Overall, it can contribute to government and public planning, as well as relevant sectors including retail, financial services, mobility, health, policing, and utilities, i.e., smart city applications, within a data-rich urban environment.
- *Smart Villages or Rural Data Science:* Rural areas or countryside are the opposite of urban areas, that include villages, hamlets, or agricultural areas. The field of Rural Data Science typically focuses on making better decisions and providing more effective solutions that include protecting public safety, providing critical health services, agriculture, and fostering economic development from a data-driven perspective, through extracting knowledge and actionable insights from the collected rural data. Advanced analytics of rural data including machine learning modeling can facilitate providing new opportunities for them to build insights and capacity to meet current needs and prepare for their futures. For instance, machine learning modeling [105] can help farmers to enhance their decisions to adopt sustainable agriculture utilizing the increasing amount of data captured by emerging technologies, e.g., the Internet of Things (IoT), mobile technologies and devices, etc. [51] [52] [4]. Overall, rural data science can play a very important role in the economic and social development of rural areas, through agriculture, business, self-employment, construction, banking, healthcare, governance, or other services, etc. that lead to smarter villages.

Overall, we can conclude that data science modeling can be used to help drive changes and improvements in almost every sector in our real-world life, where the relevant data is available to analyze. To gather the right data and extract useful knowledge or actionable insights from the data for making smart decisions is the key to data science modeling in any application domain. Based on our discussion on the above ten potential real-world application domains by taking into account data-driven smart computing and decision making, we can say that the prospects of data science and the role of data scientists are huge. The "Data Scientists" typically analyze information from multiple sources to better understand the data and business problems, and develop machine learning-based analytical modeling, or data-driven tools, or solutions, focused on advanced analytics, which can make today's computing process smarter and intelligent.

6 Challenges and Research Directions

Our study on data science and analytics, particularly data science modeling in Section 3, advanced analytics methods and smart computing in Section 4, and real-world application areas in Section 5 open several research issues in the area of data-driven business solutions and eventual data products. Thus, in this section, we summarize and discuss the challenges faced and the potential research opportunities and future directions to build data-driven products.

- Understanding the real-world business problems and associated data including nature, e.g., what forms, type, size, labels, etc., is the first challenge in the data science modeling, discussed briefly in Section

3. This is actually to identify, specify, represent and quantify the domain-specific business problems and data according to the requirements. For a data-driven effective business solution, there must be a well-defined workflow before beginning the actual data analysis work. Furthermore, gathering business data is difficult because data sources can be numerous and dynamic. As a result, gathering different forms of real-world data, such as structured, or unstructured, related to a specific business issue with legal access, which varies from application to application, is challenging. Thus, the primary task is to conduct a more in-depth analysis of data collection methods. Therefore, understanding the business problem, as well as integrating and managing the raw data gathered for efficient data analysis, may be one of the most challenging aspects of working in the field of data science and analytics.

- The next challenge is the extraction of the relevant and accurate information from the collected data mentioned above. The main focus of data scientists is typically to disclose, describe, represent, and capture data-driven intelligence for actionable insights from data. However, the real-world data may contain many ambiguous values, missing values, outliers, and meaningless data [101]. The advanced analytics methods including machine and deep learning modeling, discussed in Section 4, highly impact the quality, and availability of the data. Thus understanding real-world business scenario and associated data, to whether, how, and why they are insufficient, missing, or problematic, then extend or redevelop the existing methods, such as large-scale hypothesis testing, learning inconsistency, and uncertainty, etc. to address the complexities in data and business problems is important. Therefore, to accurately clean and pre-process the diverse data collected from multiple sources, and to prepare data for deeper analysis in the area could be another challenging task.
- Understanding and selecting the appropriate advanced analytical methods to extract the useful insights for smart decision-making for a particular business problem is the main issue in the area of data science. The emphasis of advanced analytics is more on anticipating the use of data to detect patterns to determine what is likely to occur in the future. Basic analytics offer a description of data in general, while advanced analytics is a step forward in offering a deeper understanding of data and helping to granular data analysis. Thus, understanding the advanced analytics methods, especially machine, and deep learning-based modeling is the key. The traditional learning techniques mentioned in Section 4 may not be directly applicable for the expected outcome in many cases. For instance, in a rule-based system, the traditional association rule learning technique [7] may extract redundant generation from the data that makes the decision-making process complex and ineffective [113]. Thus, a scientific understanding of the learning algorithms, mathematical properties, how the techniques are robust or fragile to input data, is needed to understand. Therefore, a deeper understanding of the strengths and drawbacks of the existing learning methods [38] [105] to solve a particular business problem is needed, consequently to improve or optimize the learning algorithms, or to propose the new techniques with higher accuracy becomes a challenging issue.
- The traditional data-driven models or systems typically use a large amount of business data to generate data-driven decisions. In several application fields, however, the new trends are more likely to be interesting and useful for modeling and predicting the future than older ones. For example, smartphone user behavior modeling, IoT services, stock market forecasting, health or transport service, job market analysis, and other related areas where time-series and actual human interests or preferences are involved over time. Thus, rather than considering the traditional data analysis, the concept of RecencyMiner, i.e., recent pattern-based extracted insight or knowledge proposed in our earlier paper Sarker et al. [108] might be effective. Therefore, to propose the new techniques by taking into account the recent data patterns, and consequently to build a recency-based data-driven model for solving real-world problems, is another challenging issue in the area.
- The most crucial task for a data-driven smart system is to create a framework that supports data science modeling discussed in Section 3. As a result, advanced analytical methods based on machine learning or deep learning techniques can be considered in such a system to make the framework capable of resolving the issues. Besides, incorporating contextual information such as temporal context, spatial context, social context, environmental context, etc. [100] can be used for building an adaptive, context-aware, and dynamic model or framework, depending on the problem domain. As a result, a well-designed framework for a specific problem domain, as well as experimental evaluation, is a very important direction, as well as a big challenge.
- In several important application areas such as autonomous cars, criminal justice, health care, recruitment, housing, management of the human resource,

public safety, where decisions made by models, or AI agents have a direct effect on human lives. As a result, there is growing concerned about whether these decisions can be trusted to be right, reasonable, ethical, personalized, accurate, robust, and secure, particularly in the context of adversarial attacks [104]. If we can explain the result in a meaningful way, then the model can be better trusted by the end-user. For machine-learned models, new trust properties yield new trade-offs, such as privacy versus accuracy; robustness versus efficiency; fairness versus robustness. Therefore, incorporating trustworthy AI particularly, data-driven or machine learning modeling could be another significant challenge in the area.

In the above, we have summarized and discussed several challenges and the potential research opportunities and directions, within the scope of our study in the area of data science and advanced analytics. The data scientists and the researchers in the relevant area have the opportunity to contribute to each issue identified above and build effective data-driven models, to make smart decisions in the corresponding business problems.

7 Conclusion

In this paper, we have presented a comprehensive view on data Science including various types of advanced analytical methods that can be applied to enhance the intelligence and the capabilities of an application. We have also visualized the current popularity of data science and machine learning-based advanced analytical modeling and also differentiate these from the relevant terms used in the area, to make the position of this paper. A thorough study on the data science modeling with its various processing layers that are needed to extract the actionable insights from the data for a particular business problem and the eventual data product. Thus, according to our goal, we have briefly discussed how different data modules can play a significant role in a data-driven business solution through the data science process. For this, we have also summarized various types of advanced analytical methods and outcomes as well as machine learning modeling that are needed to solve the associated business problems. Thus, this study's key contribution has been identified as the explanation of different advanced analytical methods and their applicability in various real-world data-driven applications areas including business, healthcare, cybersecurity, urban and rural data science, and so on by taking into account data-driven smart computing and decision making.

Finally, within the scope of our study, we outlined and discussed the challenges we faced, as well as possible research opportunities and future directions. As a result, the challenges identified provide promising research opportunities in the field that can be explored with effective solutions to improve the data-driven model and systems. Overall, we conclude that our study of advanced analytical solutions based on data science and machine learning leads in a positive direction and can be used as a reference guide for future research and applications in the field of data science and its real-world applications by both academia and industry professionals.

Compliance with ethical standards

Conflict of interest The author declares no conflict of interest.

References

1. Canadian institute of cybersecurity, university of new brunswick, iscx dataset, url http://www.unb.ca/cic/datasets/index.html/ (accessed on 20 october 2019).
2. Cic-ddos2019 [online]. available: https://www.unb.ca/cic/datasets/ddos-2019.html/ (accessed on 28 march 2020).
3. Google trends. In *https://trends.google.com/trends/*, 2019.
4. Nadia Adnan, Shahrina Md Nordin, Imran Rahman, and Amir Noor. The effects of knowledge transfer on farmers decision making toward sustainable agriculture practices. *World Journal of Science, Technology and Sustainable Development*, 2018.
5. Rakesh Agrawal, Johannes Gehrke, Dimitrios Gunopulos, and Prabhakar Raghavan. Automatic subspace clustering of high dimensional data for data mining applications. In *Proceedings of the 1998 ACM SIGMOD international conference on Management of data*, pages 94–105, 1998.
6. Rakesh Agrawal, Tomasz Imieliński, and Arun Swami. Mining association rules between sets of items in large databases. In *ACM SIGMOD Record*, volume 22, pages 207–216. ACM, 1993.
7. Rakesh Agrawal and Ramakrishnan Srikant. Fast algorithms for mining association rules. In *Proceedings of the International Joint Conference on Very Large Data Bases, Santiago Chile, pp.∼487–499.*, volume 1215, 1994.
8. David W Aha, Dennis Kibler, and Marc K Albert. Instance-based learning algorithms. *Machine learning*, 6(1):37–66, 1991.
9. Abdulrahman Al-Abassi, Hadis Karimipour, Hamed HaddadPajouh, Ali Dehghantanha, and Reza M Parizi. Industrial big data analytics: challenges and opportunities. In *Handbook of Big Data Privacy*, pages 37–61. Springer, 2020.

10. Mohammed Ali Al-Garadi, Amr Mohamed, Abdulla Khalid Al-Ali, Xiaojiang Du, Ihsan Ali, and Mohsen Guizani. A survey of machine and deep learning methods for internet of things (iot) security. *IEEE Communications Surveys & Tutorials*, 22(3):1646–1685, 2020.
11. Mihael Ankerst, Markus M Breunig, Hans-Peter Kriegel, and Jörg Sander. Optics: ordering points to identify the clustering structure. *ACM Sigmod record*, 28(2):49–60, 1999.
12. Luigi Atzori, Antonio Iera, and Giacomo Morabito. The internet of things: A survey. *Computer networks*, 54(15):2787–2805, 2010.
13. Pierre Baldi. Autoencoders, unsupervised learning, and deep architectures. In *Proceedings of ICML workshop on unsupervised and transfer learning*, pages 37–49, 2012.
14. Fabrizio Balducci, Donato Impedovo, and Giuseppe Pirlo. Machine learning applications on agricultural datasets for smart farm enhancement. *Machines*, 6(3):38, 2018.
15. George EP Box, Gwilym M Jenkins, Gregory C Reinsel, and Greta M Ljung. *Time series analysis: forecasting and control*. John Wiley & Sons, 2015.
16. Leo Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.
17. Malte Brettel, Niklas Friederichsen, Michael Keller, and Marius Rosenberg. How virtualization, decentralization and network building change the manufacturing landscape: an industry 4.0 perspective. *FormaMente*, 12, 2017.
18. Huanhuan Cao, Tengfei Bao, Qiang Yang, Enhong Chen, and Jilei Tian. An effective approach for mining mobile user habits. In *Proceedings of the International Conference on Information and knowledge management, Toronto, ON, Canada, 26-30 October, pp.∼1677–1680. ACM, New York, USA*, 2010.
19. Longbing Cao. Data science: a comprehensive overview. *ACM Computing Surveys (CSUR)*, 50(3):1–42, 2017.
20. Gail A Carpenter and Stephen Grossberg. A massively parallel architecture for a self-organizing neural pattern recognition machine. *Computer vision, graphics, and image processing*, 37(1):54–115, 1987.
21. H Frank Cervone. Informatics and data science: an overview for the information professional. *Digital Library Perspectives*, 2016.
22. Anatole Chessel. An overview of data science uses in bioimage informatics. *Methods*, 115:110–118, 2017.
23. François Chollet. Xception: Deep learning with depthwise separable convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1251–1258, 2017.
24. Robert Cudeck. Exploratory factor analysis. In *Handbook of applied multivariate statistics and mathematical modeling*, pages 265–296. Elsevier, 2000.
25. Amitabha Das, Wee-Keong Ng, and Yew-Kwong Woon. Rapid association rule mining. In *Proceedings of the tenth international conference on Information and knowledge management*, pages 474–481. ACM, 2001.
26. Renato Cordeiro de Amorim. Constrained clustering with minkowski weighted k-means. In *2012 IEEE 13th International Symposium on Computational Intelligence and Informatics (CINTI)*, pages 13–17. IEEE, 2012.
27. Himel Dev and Zhicheng Liu. Identifying frequent user tasks from application logs. In *Proceedings of the 22nd International Conference on Intelligent User Interfaces*, pages 263–273, 2017.
28. David Donoho. 50 years of data science. *Journal of Computational and Graphical Statistics*, 26(4):745–766, 2017.
29. Nathan Eagle and Alex Sandy Pentland. Reality mining: sensing complex social systems. *Personal and ubiquitous computing*, 10(4):255–268, 2006.
30. Zeynep Engin, Justin van Dijk, Tian Lan, Paul A Longley, Philip Treleaven, Michael Batty, and Alan Penn. Data-driven urban management: Mapping the landscape. *Journal of Urban Management*, 9(2):140–150, 2020.
31. Martin Ester, Hans-Peter Kriegel, Jörg Sander, Xiaowei Xu, et al. A density-based algorithm for discovering clusters in large spatial databases with noise. In *Kdd*, volume 96, pages 226–231, 1996.
32. Peter A Flach and Nicolas Lachiche. Confirmation-guided discovery of first-order rules with tertius. *Machine Learning*, 42(1-2):61–95, 2001.
33. Yoav Freund, Robert E Schapire, et al. Experiments with a new boosting algorithm. In *Icml*, volume 96, pages 148–156. Citeseer, 1996.
34. Prajakta Ghavare and Prashant Ahire. Big data classification of users navigation and behavior using web server logs. In *2018 Fourth International Conference on Computing Communication Control and Automation (ICCUBEA)*, pages 1–6. IEEE, 2018.
35. Ian Goodfellow, Yoshua Bengio, Aaron Courville, and Yoshua Bengio. *Deep learning*, volume 1. MIT press Cambridge, 2016.
36. Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014.
37. Martin Halvey, Mark T Keane, and Barry Smyth. Time based segmentation of log data for user navigation prediction in personalization. In *Proceedings of the International Conference on Web Intelligence, Compiegne, France, 19-22 September, pp.∼636–640. IEEE Computer Society, Washington, DC, USA.*, 2005.
38. Jiawei Han, Jian Pei, and Micheline Kamber. *Data mining: concepts and techniques*. Elsevier, Amsterdam, Netherlands, 2011.
39. Jiawei Han, Jian Pei, and Yiwen Yin. Mining frequent patterns without candidate generation. In *ACM Sigmod Record*, volume 29, pages 1–12. ACM, 2000.
40. Seng Hansun. A new approach of moving average method in time series analysis. In *2013 conference on new media studies (CoNMedia)*, pages 1–4. IEEE, 2013.
41. Stephanie A Harmon, Thomas H Sanford, Sheng Xu, Evrim B Turkbey, Holger Roth, Ziyue Xu, Dong Yang, Andriy Myronenko, Victoria Anderson, Amel Amalou, et al. Artificial intelligence for the detection of covid-19 pneumonia on chest ct using multinational datasets. *Nature communications*, 11(1):1–7, 2020.
42. Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Spatial pyramid pooling in deep convolutional networks for visual recognition. *IEEE transactions on pattern analysis and machine intelligence*, 37(9):1904–1916, 2015.
43. Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

44. Pinjia He, Jieming Zhu, Shilin He, Jian Li, and Michael R Lyu. Towards automated log parsing for large-scale log data analysis. *IEEE Transactions on Dependable and Secure Computing*, 15(6):931–944, 2017.
45. Fatemeh Hemmatian and Mohammad Karim Sohrabi. A survey on classification techniques for opinion mining and sentiment analysis. *Artificial Intelligence Review*, pages 1–51, 2019.
46. Geoffrey E Hinton. A practical guide to training restricted boltzmann machines. In *Neural networks: Tricks of the trade*, pages 599–619. Springer, 2012.
47. Maurice Houtsma and Arun Swami. Set-oriented mining for association rules in relational databases. In *Data Engineering, 1995. Proceedings of the Eleventh International Conference on*, pages 25–33. IEEE, 1995.
48. Matt C Howard. A review of exploratory factor analysis decisions and overview of current practices: What we are doing and how can we improve? *International Journal of Human-Computer Interaction*, 32(1):51–62, 2016.
49. George H John and Pat Langley. Estimating continuous distributions in bayesian classifiers. In *Proceedings of the Eleventh conference on Uncertainty in artificial intelligence*, pages 338–345. Morgan Kaufmann Publishers Inc., 1995.
50. Emilia Kacprzak, Laura Koesten, Luis-Daniel Ibáñez, Tom Blount, Jeni Tennison, and Elena Simperl. Characterising dataset search—an analysis of search logs and data requests. *Journal of Web Semantics*, 55:37–55, 2019.
51. Sachin S Kamble, Angappa Gunasekaran, and Shradha A Gawankar. Sustainable industry 4.0 framework: A systematic literature review identifying the current trends and future perspectives. *Process Safety and Environmental Protection*, 117:408–425, 2018.
52. Sachin S Kamble, Angappa Gunasekaran, and Shradha A Gawankar. Achieving sustainable performance in a data-driven agriculture supply chain: A review for research and applications. *International Journal of Production Economics*, 219:179–194, 2020.
53. Anuj Karpatne, Gowtham Atluri, James H Faghmous, Michael Steinbach, Arindam Banerjee, Auroop Ganguly, Shashi Shekhar, Nagiza Samatova, and Vipin Kumar. Theory-guided data science: A new paradigm for scientific discovery from data. *IEEE Transactions on knowledge and data engineering*, 29(10):2318–2331, 2017.
54. Leonard Kaufman and Peter J Rousseeuw. *Finding groups in data: an introduction to cluster analysis*, volume 344. John Wiley & Sons, 2009.
55. S. Sathiya Keerthi, Shirish Krishnaj Shevade, Chiranjib Bhattacharyya, and Karuturi Radha Krishna Murthy. Improvements to platt's smo algorithm for svm classifier design. *Neural computation*, 13(3):637–649, 2001.
56. Vijay Khadse, Parikshit N Mahalle, and Swapnil V Biraris. An empirical comparison of supervised machine learning algorithms for internet of things data. In *2018 Fourth International Conference on Computing Communication Control and Automation (ICCUBEA)*, pages 1–6. IEEE, 2018.
57. Tatsuaki Kimura, Akio Watanabe, Tsuyoshi Toyono, and Keisuke Ishibashi. Proactive failure detection learning generation patterns of large-scale network logs. *IEICE Transactions on Communications*, 2018.
58. Teuvo Kohonen. The self-organizing map. *Proceedings of the IEEE*, 78(9):1464–1480, 1990.
59. Nickolaos Koroniotis, Nour Moustafa, Elena Sitnikova, and Benjamin Turnbull. Towards the development of realistic botnet dataset in the internet of things for network forensic analytics: Bot-iot dataset. *Future Generation Computer Systems*, 100:779–796, 2019.
60. Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.
61. Dmytro Krukovets et al. Data science opportunities at central banks: Overview. *Visnyk of the National Bank of Ukraine*, 249:13–24, 2020.
62. Merima Kulin, Carolina Fortuna, Eli De Poorter, Dirk Deschrijver, and Ingrid Moerman. Data-driven design of intelligent wireless networks: An overview and tutorial. *Sensors*, 16(6):790, 2016.
63. Donghwoon Kwon, Hyunjoo Kim, Jinoh Kim, Sang C Suh, Ikkyun Kim, and Kuinam J Kim. A survey of deep learning-based network anomaly detection. *Cluster Computing*, 22(1):949–961, 2019.
64. Prasanth Lade, Rumi Ghosh, and Soundar Srinivasan. Manufacturing analytics and industrial internet of things. *IEEE Intelligent Systems*, 32(3):74–79, 2017.
65. Deanne Larson and Victor Chang. A review and future direction of agile, business intelligence, analytics and data science. *International Journal of Information Management*, 36(5):700–710, 2016.
66. Saskia Le Cessie and Johannes C Van Houwelingen. Ridge estimators in logistic regression. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 41(1):191–201, 1992.
67. Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.
68. Jay Lee, Behrad Bagheri, and Hung-An Kao. Recent advances and trends of cyber-physical systems and big data analytics in industrial informatics. In *International proceeding of int conference on industrial informatics (INDIN)*, pages 1–6, 2014.
69. Joffrey L Leevy, Taghi M Khoshgoftaar, Richard A Bauder, and Naeem Seliya. A survey on addressing high-class imbalance in big data. *Journal of Big Data*, 5(1):42, 2018.
70. Zuhe Li, Yangyu Fan, Bin Jiang, Tao Lei, and Weihua Liu. A survey on sentiment analysis and opinion mining for social multimedia. *Multimedia Tools and Applications*, 78(6):6939–6967, 2019.
71. Bing Liu. *Sentiment analysis: Mining opinions, sentiments, and emotions*. Cambridge university press, 2020.
72. Jiaying Liu, Tao Tang, Wei Wang, Bo Xu, Xiangjie Kong, and Feng Xia. A survey of scholarly data visualization. *Ieee Access*, 6:19205–19221, 2018.
73. Bing Liu Wynne Hsu Yiming Ma. Integrating classification and association rule mining. In *Proceedings of the fourth international conference on knowledge discovery and data mining*, 1998.
74. Chuang Ma, Hao Helen Zhang, and Xiangfeng Wang. Machine learning for big data analytics in plants. *Trends in plant science*, 19(12):798–808, 2014.
75. James MacQueen et al. Some methods for classification and analysis of multivariate observations. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, volume 1, pages 281–297. Oakland, CA, USA, 1967.
76. André Marchand and Paul Marx. Automated product recommendations with preference-based explanations. *Journal of retailing*, 96(3):328–343, 2020.

77. Abhinav Mehrotra, Robert Hendley, and Mirco Musolesi. Prefminer: mining user's preferences for intelligent mobile notification management. In *Proceedings of the International Joint Conference on Pervasive and Ubiquitous Computing, Heidelberg, Germany, 12-16 September, pp.∼1223–1234. ACM, New York, USA.*, 2016.
78. Youssoufa Mohamadou, Aminou Halidou, and Pascalin Tiam Kapen. A review of mathematical modeling, artificial intelligence and datasets used in the study, prediction and management of covid-19. *Applied Intelligence*, 50(11):3913–3925, 2020.
79. Nour Moustafa and Jill Slay. Unsw-nb15: a comprehensive data set for network intrusion detection systems (unsw-nb15 network data set). In *2015 military communications and information systems conference (MilCIS)*, pages 1–6. IEEE, 2015.
80. U Nations. Revision of world urbanization prospects. *United Nations: New York, NY, USA*, 2018.
81. Mehrbakhsh Nilashi, Othman bin Ibrahim, Hossein Ahmadi, and Leila Shahmoradi. An analytical method for diseases prediction using machine learning techniques. *Computers & Chemical Engineering*, 106:212–223, 2017.
82. Worapat Paireekreng, Kowit Rapeepisarn, and Kok Wai Wong. Time-based personalised mobile game downloading. In *Transactions on Edutainment II, pp.∼59–69.* 2009.
83. Yue Pan, Limao Zhang, and Zhiwu Li. Mining event logs for knowledge discovery based on adaptive efficient fuzzy kohonen clustering network. *Knowledge-Based Systems*, 209:106482, 2020.
84. Hae-Sang Park and Chi-Hyuck Jun. A simple and fast algorithm for k-medoids clustering. *Expert systems with applications*, 36(2):3336–3341, 2009.
85. Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, 12:2825–2830, 2011.
86. Sajida Perveen, Muhammad Shahbaz, Karim Keshavjee, and Aziz Guergachi. Metabolic syndrome and development of diabetes mellitus: Predictive modeling based on machine learning techniques. *IEEE Access*, 7:1365–1375, 2018.
87. Gabriel Peyré, Marco Cuturi, et al. Computational optimal transport: With applications to data science. *Foundations and Trends® in Machine Learning*, 11(5-6):355–607, 2019.
88. Santi Phithakkitnukoon, Ram Dantu, Rob Claxton, and Nathan Eagle. Behavior-based adaptive call predictor. *ACM Transactions on Autonomous and Adaptive Systems*, 6(3):21:1–21:28), 2011.
89. Samira Pouyanfar, Yimin Yang, Shu-Ching Chen, Mei-Ling Shyu, and SS Iyengar. Multimedia big data analytics: A survey. *ACM computing surveys (CSUR)*, 51(1):1–34, 2018.
90. Foster Provost and Tom Fawcett. *Data Science for Business: What you need to know about data mining and data-analytic thinking.* " O'Reilly Media, Inc.", 2013.
91. Xuedi Qin, Yuyu Luo, Nan Tang, and Guoliang Li. Making data visualization more efficient and effective: a survey. *The VLDB Journal*, 29(1):93–117, 2020.
92. J. Ross Quinlan. Induction of decision trees. *Machine learning*, 1(1):81–106, 1986.
93. J. Ross Quinlan. C4.5: Programs for machine learning. *Machine Learning*, 1993.
94. Carl Rasmussen. The infinite gaussian mixture model. *Advances in neural information processing systems*, 12:554–560, 1999.
95. Reza Rawassizadeh, Martin Tomitsch, Katarzyna Wac, and A Min Tjoa. Ubiqlog: a generic mobile phone-based life-log framework. *Personal and ubiquitous computing*, 17(4):621–637, 2013.
96. Bernd Resch and Michael Szell. Human-centric data science for urban studies, 2019.
97. Aya Rizk and Ahmed Elragal. Data science: developing theoretical contributions in information systems via text analytics. *Journal of Big Data*, 7(1):1–26, 2020.
98. Lior Rokach. A survey of clustering algorithms. In *Data Mining and Knowledge Discovery Handbook*, pages 269–298. Springer, 2010.
99. Saima Safdar, Saad Zafar, Nadeem Zafar, and Naurin Farooq Khan. Machine learning based decision support systems (dss) for heart disease diagnosis: a review. *Artificial Intelligence Review*, 50(4):597–623, 2018.
100. Iqbal H Sarker. Context-aware rule learning from smartphone data: survey, challenges and future directions. *Journal of Big Data*, 6(1):1–25, 2019.
101. Iqbal H Sarker. A machine learning based robust prediction model for real-life mobile phone data. *Internet of Things*, 5:180–193, 2019.
102. Iqbal H Sarker. Ai-driven cybersecurity: An overview, security intelligence modeling and research directions. *SN Computer Science*, 2021.
103. Iqbal H Sarker. Cyberlearning: Effectiveness analysis of machine learning security modeling to detect cyber-anomalies and multi-attacks. *Internet of Things*, page 100393, 2021.
104. Iqbal H Sarker. Deep cybersecurity: A comprehensive overview from neural network and deep learning perspective. *SN Computer Science*, 2021.
105. Iqbal H Sarker. Machine learning: Algorithms, real-world applications and research directions. *SN Computer Science*, 2(3):1–21, 2021.
106. Iqbal H Sarker, Yoosef B Abushark, Fawaz Alsolami, and Asif Irshad Khan. Intrudtree: A machine learning based cyber security intrusion detection model. *Symmetry*, 12(5):754, 2020.
107. Iqbal H Sarker, Hamed Alqahtani, Fawaz Alsolami, Asif Irshad Khan, Yoosef B Abushark, and Mohammad Khubeb Siddiqui. Context pre-modeling: an empirical analysis for classification based user-centric context-aware predictive modeling. *Journal of Big Data*, 7(1):1–23, 2020.
108. Iqbal H Sarker, Alan Colman, and Jun Han. Recencyminer: mining recency-based personalized behavior from contextual smartphone data. *Journal of Big Data*, 6(1):1–21, 2019.
109. Iqbal H Sarker, Alan Colman, Jun Han, Asif Irshad Khan, Yoosef B Abushark, and Khaled Salah. Behavdt: a behavioral decision tree learning to build user-centric context-aware predictive model. *Mobile Networks and Applications*, 25(3):1151–1161, 2020.
110. Iqbal H Sarker, Alan Colman, Muhammad Ashad Kabir, and Jun Han. Phone call log as a context source to modeling individual user behavior. In *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing (Ubicomp): Adjunct, Germany*, pages 630–634. ACM, 2016.
111. Iqbal H Sarker, Alan Colman, Muhammad Ashad Kabir, and Jun Han. Individualized time-series segmentation

for mining mobile phone user behavior. *The Computer Journal*, 61(3):349–368, 2018.

112. Iqbal H Sarker, Mohammed Moshiul Hoque, Md Kafil Uddin, and Tawfeeq Alsanoosy. Mobile data science and intelligent apps: Concepts, ai-based modeling and research directions. *Mobile Networks and Applications*, pages 1–19, 2020.
113. Iqbal H Sarker and ASM Kayes. Abc-ruleminer: User behavioral rule-based machine learning method for context-aware intelligent services. *Journal of Network and Computer Applications*, page 102762, 2020.
114. Iqbal H Sarker, ASM Kayes, Shahriar Badsha, Hamed Alqahtani, Paul Watters, and Alex Ng. Cybersecurity data science: an overview from machine learning perspective. *Journal of Big Data*, 7(1):1–29, 2020.
115. Iqbal H Sarker, ASM Kayes, and Paul Watters. Effectiveness analysis of machine learning classification models for predicting personalized context-aware smartphone usage. *Journal of Big Data*, 6(1):1–28, 2019.
116. Markus Schläpfer, Luís MA Bettencourt, Sébastian Grauwin, Mathias Raschke, Rob Claxton, Zbigniew Smoreda, Geoffrey B West, and Carlo Ratti. The scaling of human interactions with city size. *Journal of the Royal Society Interface*, 11(98):20130789, 2014.
117. Nishant Shukla and Kenneth Fricklas. *Machine learning with TensorFlow*. Manning Greenwich, 2018.
118. Sima Siami-Namini, Neda Tavakoli, and Akbar Siami Namin. A comparison of arima and lstm in forecasting time series. In *2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA)*, pages 1394–1401. IEEE, 2018.
119. Gökhan Silahtaroğlu and Nevin Yılmaztürk. Data analysis in health and big data: A machine learning medical diagnosis model based on patients' complaints. *Communications in Statistics-Theory and Methods*, pages 1–10, 2019.
120. Andrea Silvestrini and David Veredas. Temporal aggregation of univariate and multivariate time series models: a survey. *Journal of Economic Surveys*, 22(3):458–497, 2008.
121. Beata Ślusarczyk. Industry 4.0: Are we ready? *Polish Journal of Management Studies*, 17, 2018.
122. Peter HA Sneath. The application of computers to taxonomy. *Journal of General Microbiology*, 17(1), 1957.
123. Thorvald Sorensen. method of establishing groups of equal amplitude in plant sociology based on similarity of species. *Biol. Skr.*, 5, 1948.
124. Vijay Srinivasan, Saeed Moghaddam, and Abhishek Mukherji. Mobileminer: Mining your frequent patterns on your phone. In *Proceedings of the International Joint Conference on Pervasive and Ubiquitous Computing, Seattle, WA, USA, 13-17 September, pp.∼389–400. ACM, New York, USA.*, 2014.
125. Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–9, 2015.
126. Arman Tajbakhsh, Mohammad Rahmati, and Abdolreza Mirzaei. Intrusion detection using fuzzy association rules. *Applied Soft Computing*, 9(2):462–469, 2009.
127. Mahbod Tavallaee, Ebrahim Bagheri, Wei Lu, and Ali A Ghorbani. A detailed analysis of the kdd cup 99 data set. In *2009 IEEE Symposium on Computational Intelligence for Security and Defense Applications*, pages 1–6. IEEE, 2009.
128. Manos Tsagkias, Tracy Holloway King, Surya Kallumadi, Vanessa Murdock, and Maarten de Rijke. Challenges and research opportunities in ecommerce search and recommendations. In *ACM SIGIR Forum*, volume 54, pages 1–23. ACM New York, NY, USA, 2021.
129. Chun-Wei Tsai, Chin-Feng Lai, Han-Chieh Chao, and Athanasios V Vasilakos. Big data analytics: a survey. *Journal of Big data*, 2(1):1–32, 2015.
130. Kerem Sinan Tuncel and Mustafa Gokce Baydogan. Autoregressive forests for multivariate time series modeling. *Pattern recognition*, 73:202–215, 2018.
131. Kiri Wagstaff, Claire Cardie, Seth Rogers, Stefan Schrödl, et al. Constrained k-means clustering with background knowledge. In *Icml*, volume 1, pages 577–584, 2001.
132. JunPing Wang, WenSheng Zhang, YouKang Shi, ShiHui Duan, and Jin Liu. Industrial big data analytics: challenges, methodologies, and applications. *arXiv preprint arXiv:1807.01016*, 2018.
133. Liye Wang, Jin Zhang, Guoqing Chen, and Dandan Qiao. Identifying comparable entities with indirectly associative relations and word embeddings from web search logs. *Decision Support Systems*, 141:113465, 2021.
134. Wei Wang, Jiong Yang, Richard Muntz, et al. Sting: A statistical information grid approach to spatial data mining. In *VLDB*, volume 97, pages 186–195, 1997.
135. Michael L Waskom. Seaborn: statistical data visualization. *Journal of Open Source Software*, 6(60):3021, 2021.
136. Peng Wei, Yufeng Li, Zhen Zhang, Tao Hu, Ziyong Li, and Diyang Liu. An optimization method for intrusion detection classification model based on deep belief network. *IEEE Access*, 7:87593–87605, 2019.
137. Karl Weiss, Taghi M Khoshgoftaar, and DingDing Wang. A survey of transfer learning. *Journal of Big data*, 3(1):9, 2016.
138. Ian H Witten and Eibe Frank. *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann, 2005.
139. Ian H Witten, Eibe Frank, Leonard E Trigg, Mark A Hall, Geoffrey Holmes, and Sally Jo Cunningham. Weka: Practical machine learning tools and techniques with java implementations. 1999.
140. Yang Xin, Lingshuang Kong, Zhi Liu, Yuling Chen, Yanmiao Li, Hongliang Zhu, Mingcheng Gao, Haixia Hou, and Chunhua Wang. Machine learning and deep learning methods for cybersecurity. *IEEE Access*, 6:35365–35381, 2018.
141. Dongkuan Xu and Yingjie Tian. A comprehensive survey of clustering algorithms. *Annals of Data Science*, 2(2):165–193, 2015.
142. Jing Ya, Tingwen Liu, Quangang Li, Jinqiao Shi, Haoliang Zhang, Pin Lv, and Li Guo. Mining host behavior patterns from massive network and security logs. *Procedia Computer Science*, 108:38–47, 2017.
143. An Gie Yong, Sean Pearce, et al. A beginner's guide to factor analysis: Focusing on exploratory factor analysis. *Tutorials in quantitative methods for psychology*, 9(2):79–94, 2013.
144. Mohammed Javeed Zaki. Scalable algorithms for association mining. *IEEE transactions on knowledge and data engineering*, 12(3):372–390, 2000.
145. Qiankun Zhao and Sourav S Bhowmick. Association rule mining: A survey. *Nanyang Technological University, Singapore*, 2003.

146. Pei Zheng and Lionel M Ni. Spotlight: the rise of the smart phone. *IEEE Distributed Systems Online*, 7(3):3–3, 2006.
147. Tao Zheng, Wei Xie, Liling Xu, Xiaoying He, Ya Zhang, Mingrong You, Gong Yang, and You Chen. A machine learning-based framework to identify type 2 diabetes through electronic health records. *International journal of medical informatics*, 97:120–127, 2017.
148. Zhi-Jie Zhou, Guan-Yu Hu, Chang-Hua Hu, Cheng-Lin Wen, and Lei-Lei Chang. A survey of belief rule-base expert system. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 2019.
149. Hengshu Zhu, Enhong Chen, Hui Xiong, Kuifei Yu, Huanhuan Cao, and Jilei Tian. Mining mobile user preferences for personalized context-aware recommendation. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 5(4):58, 2014.
150. He Zikang, Yang Yong, Yang Guofeng, and Zhang Xinyu. Sentiment analysis of agricultural product ecommerce review data based on deep learning. In *2020 International Conference on Internet of Things and Intelligent Applications (ITIA)*, pages 1–7. IEEE, 2020.