https://everant.in/index.php/mcsj

Research Article

Mathematics and Computer Science Journal

# An Overview of Data Science Algorithms

**Vishwanadham Mandala**

Service Delivery Lead, Cummins Inc

ARTICLE INFO

ABSTRACT

**Vishwanadham Mandala,**
**Service Delivery Lead,**
**Cummins Inc,**

**Abstract**

Data science algorithms are on the way to becoming an integral part of every company, and we can already see the effects in many corporations that have invented their own data science teams and also implemented the latest data science algorithms. To be able to work with all the different challenges that are emerging, new powerful data science tools have been developed (e.g. Python, R, H2O, Weka, Tensorflow, Spark, Flink, BigML or KNIME). One balance that companies that want to use these new tools have to face is the cost of implementation vs. the enhanced development that they give in return. Nowadays, most of the advanced algorithms are open source and available on multiple platforms and programming languages, which helps to minimize the development cost challenges that each company has to overcome.

Still, one of the main dangers lurking inside these development teams is that they do not know what the state of the art of advanced algorithms is and which problem they can address. To help mitigate this problem, a review of algorithms has been implemented in this paper. This review gives us a perspective on which algorithms are being developed and which problem areas they can address. With the development of more powerful data science algorithms, we are also enabling the possibility of tackling more complex and interesting problems. However, one characteristic of the review is that there are missing algorithms from the many that are currently being produced and frequently selected by the community as the best performers in many benchmark datasets.

**Keywords:** Data Science Algorithms, Industry 4.0, Internet of Things (IoT), Artificial Intelligence (AI), Machine Learning (ML), Smart Manufacturing (SM),Computer Science, Data Science,Vehicle, Vehicle Reliability

## 1. Introduction

In the modern IoT and data-driven world, a significant amount of data is being produced every second. The speed of data creation is fast enough to make data storage appear almost insignificant. Nowadays, data creation is not only limited to

social media and technology companies. The finance sector, healthcare institutions, and factories are also producing a large amount of data. The interesting thing is that this big data will make traditional data storage and machine learning algorithms almost unacceptable in the process of data analysis and extraction of important information, and possibly so shortly, as handling a large amount of data using traditional data science techniques is not effective. It is clear that traditional machine learning techniques, no matter how robust, will not perform satisfactorily in this big data world. It is for these reasons that researchers and organizations are trying to adapt machine learning algorithms to manage such a huge amount of data. Data science, a discipline that incorporates knowledge from several areas including mathematics, statistics, and computer science, has shown significant potential in this direction. In this new discipline, which excels in data analysis and knowledge extraction, there are already several algorithms, and new ones will be developed. This paper aims to present an in-depth study of data science techniques, discussing several algorithms studied, and then applying some of them in datasets that are accessible to everyone. The paper is organized as follows. In the second section, we introduce some of the key principles, timelines, and evolution of data science algorithms. In the third section, we focus on presenting the current most used algorithms in the data science field, and some new ones. In the fourth section, three types of data science algorithms are applied to well-known datasets. In the fifth section, the main conclusions will be taken. Finally, in the sixth section, the main contributions of the paper and future work will be introduced.

### 1.1. Background and Significance of Data Science Algorithms

With the rise of big data and data analytics, data science remains a very hot topic in recent years. Generally speaking, data science is the study of capturing, storing, analyzing, and communicating data. It integrates many scientific fields including algorithms, mathematics, computer science, machine learning, data mining, and so on, to extract knowledge and insights from data. As a vital component of data science, big data techniques are applied to handle the data that are too large or complex for traditional data-processing applications. Compared with other important components, such as creating and analyzing sophisticated models, producing the most accurate data can help you achieve your objectives more effectively.

To accomplish the objective of the paper, the rest of the paper is organized as follows. In Sec. 2, a comprehensive summary of data science algorithms is introduced, which includes the major categories, fronts, and some cutting-edge ones. More specifically, machine learning, deep learning, statistics, and data mining methods form the category of data science algorithms which is introduced in this section. Since the rise of big data and data science, big data techniques have become a popular and prosperous research area. In this work, big data techniques take responsibility for handling and representing large-scale data noumenon, that is data representation and processing.


**Fig 1: Data Science Algorithms**

## 2. Common Data Science Algorithms

I encourage you to chat with your students and cross-check at the end. Say that you have any learning difficulty, tell them the content of what they say, without being discouraged. But I am here evaluating this help against what you already know, the contents of which teachers have already

circulated, the questions you have asked and will ask, and I have worked to facilitate your breakthroughs, do not forget. Coat on the seat in one, to review at the end. Shall we try and see what your fingers can creak now?There have been many developments in data science algorithms in 2020. In addition to popular algorithms, long-time forgotten algorithms have also been implemented again through different machine learning libraries. In this post, I will try to give a brief overview of my experiences. Data Science is an interesting field. It is related to extracting information from data with the help of algorithms. In this field, there are certain subfields such as data mining, machine learning, etc. and usually, these subfields are used interchangeably. While data mining is concerned with exploring and developing novel techniques to make previously inaccessible information apparent, machine learning is interested in answering a specific question from a target concept, which requires learning from a previously studied domain. A target concept is a specific function that assigns inputs to outputs, or when viewed probabilistically, a probability distribution over a domain.

### 2.1.Supervised Learning Algorithms

Computers are fantastic tools for many applications involving well-defined rules and logic, but some tasks remain challenging for machines to perform. One simple example is recognizing handwritten text: children quite rapidly master the skill without needing explicit instructions or formal rules about what to look for. Such higher-level thought activities are some of the last bastions where humans still have an edge over machines, but recent progress in several areas of Artificial Intelligence (AI) demonstrates that they are no longer unbreakable barriers to automation.

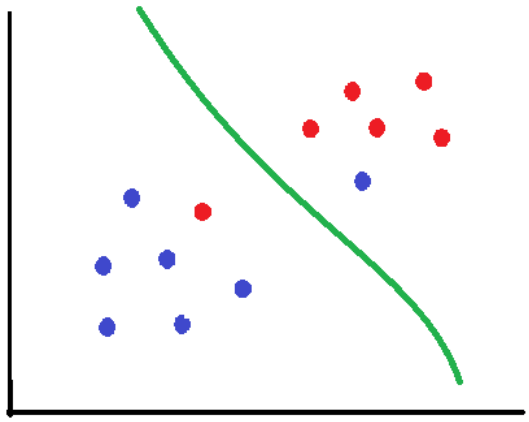

**Fig 2 : Supervised Learning**

The field of AI predates modern computer science and big data technologies by a lot. Since the beginning of the field, practitioners have been interested in teaching machines how to learn. The ability to learn and optimize some criteria (simultaneously or sequentially) is the defining property of what is currently called data science or machine learning (ML). The reason for the rebranding is due to the increase in technical complexity and ubiquity of ML.There are many algorithms that data scientists use to build this type of model. Below we will see the most commonly used ones, organized by the type of data they process and the type of problem they tend to solve most often.

### 2.2. Unsupervised Learning Algorithms

Unsupervised learning models take a table of data and learn techniques for summarizing the distribution of the data. These algorithms don't require human intervention. One important result is the extraction of clusters, which consist of observations that share some property. Clustering is necessary in various fields. For example, in retail chain marketing, it is important to find a group of customers who share the same interests to promote a product.

The classic examples of unsupervised learning are the use of clustering such as k-means clustering, hierarchical clustering, and spectral clustering,

although other popular examples include principal component analysis and density estimation. Some models, mostly those that use a probabilistic graph to describe relationships among variables, fall into both the supervised and unsupervised categories. Also, techniques in unsupervised learning can be useful for exploratory data analysis and for feature expansion (augmenting the table of data with new features).

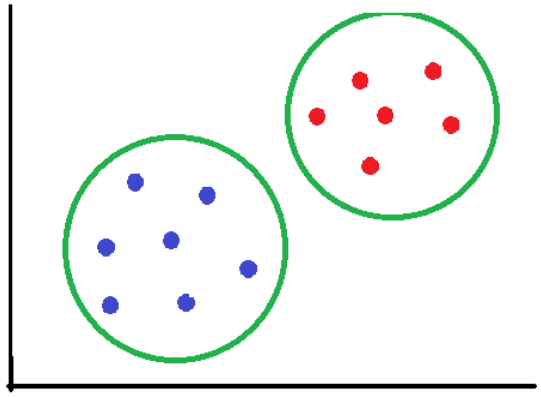

**Fig 3: unsupervised learning**

### 3. Emerging Trends and Innovations in Data Science Algorithms

Innovations in the algorithms used for predictive modeling require the use of larger datasets and more complex processing to achieve more accurate and generalizable results. As a result, these processes have acquired a level of complexity that has given rise to terms such as data science.Data science refers to the practice of applying advanced analytics to large and diverse datasets, which may include datasets of various shapes and sizes, and structured and semi-structured data, including unstructured text, images, and sound files. Starting from a dataset with labeled examples or cases, a data science pipeline can include iterative processes that progressively increase the performance of a model to achieve predictive results that are often deployed in real-world applications.Interest in AI and ML appeals to the idea that machines can learn from data. This is very appealing in scenarios where models that describe the intrinsic relationships between the variables exhibited in the input data can be found so that predictions can be formulated. An ML-based model aims to find conclusions embedded in the input data, typically labeled training data, which is then used for analysis or generating predictions.Recently, artificial intelligence (AI) and its main component, machine learning (ML), have become terms of high public visibility and discussion. They are frequently used as umbrella terms for processes that define the latest wave of software capabilities. In data science, these two terms have been for some time at the core of common practices used to design and develop models able to achieve accurate and generalizable predictive results and forecasts.Thus, while the concepts have been known and used since the 1970s, at the heart of the ML process, this can uncover unknown patterns, classify particular clusters of data, or recognize specific structures among the data. Previously known concepts include rule-based expert systems that found hypotheses or decisions based on rules explicitly codified as data if-then constructs; canonical pattern recognition techniques such as template matching; k-nearest neighbor, linear discriminant analysis, and dimensionality reduction; evolutionary symbolic algorithms that consider the evolution of an implicit mathematical model; and also Bayesian classifiers or supervised learning techniques based on inductive reasoning that infer knowledge about the distribution of unobserved events.

### 4. Applications of Data Science Algorithms in Various Industries

In this paper, an overview of the data science algorithms used for model building in 2020 is presented. The DM methodology consists of several steps with the core intellectual property being the 76 data science algorithms available for use by the users. The goal of this research paper is to demonstrate and doctrine of the current data science algorithms and to inform its interested and affected stakeholders. With the increasing applications of data science algorithms in the business, medical,

financial, etc., this review paper on data science algorithms will guide data scientists, decision-makers, and research students to the key data science algorithms. Plenty of research papers on big data analytics and related topics were published in 2019 itself and work is expected to continue more aggressively in 2020 and beyond.To start, the DM process is linked to the core intellectual property surrounding the 76 data science algorithms as shown. The system development steps around the DM process are: define the problem and define the project plan; prepare the data and data enrichment; data cleansing; data transformation; data reduction; data processes; and pattern evolution. Theory; select the method and validate the model The research process is shown in a hierarchical taxonomy representing the three research options of descriptive research, decision science research, or formal proposition research. Each of these options has a specialized categorization associated with it: descriptive research, which includes research design options, context, data collection techniques, sample, data preparation requirements, analysis options, and presentation of results. This paper provides a survey on the interconnection of two major components in the area of data storage: the concept of NoSQL databases and their implementation together with data mining modeling and its key algorithms.

## 5. Conclusion

In the past, era solutions have occurred usually in a divide-plus-rule style. Experts skilled in explicit techniques or some area of expertise go away to pursue also approach improvements or recent aspects implemented after so much science. However, the present environment wants the need for the well-being of partial individuals engaged concerning data, software and metrics to be able to whole among algorithms at once. This suggests so much that the innovative developments all over the field call for an after-stand built-in which shines all whole regarding information also the life cycle of data science in a greater holistic way.Now AI systems compete together with scientific human beings in solving problems, and there are activities involving trust or criteria on how its discipline grows. Moreover deals such as afterlooks what kind of ethics, fairness, stability, display, and extra duties call for in imitation to be met. The suggestions concerning a new or repurposed algorithmic development as such satisfies some of its criteria, especially securing human welfare, are growing currently. Likewise, the need for responsible progress or deployment of AI incarnate among the process at a close distance to yet collective tasks remains a pivotal challenge.

### 5.1. Future Trends

Recently, deep learning has attracted great interest in computer vision, speech recognition, natural language processing, audio recognition, social network filtering, machine translation, bioinformatics, drug design, and many other fields. Self-improving techniques and large datasets are the main drivers of success. Deep learning is responsible for breakthroughs in image classification, automatic translation, medical diagnostics, and more. Many important problems can be solved without using deep learning. Machine learning is not equal to deep learning. Deep learning is not equal to neural networks. There is a large interest in developing self-improving systems capable of large-scale reasoning. Such systems are expected to have a great impact on finance, military, insurance, healthcare, and cybersecurity.There are many learning theories underlying various data science algorithms and a great interest in the epistemology of machine learning. Researchers aim to understand why overparameterization works. They discuss different approaches to overfitting, search for optimal model complexity and optimal model capacity, aim to explain the bias-variance trade-off, and discuss the model selection criteria. They study generalization, create an ideal learner model, analyze deep learning, and provide upper bounds on the generalization error. They carefully

analyze statistical and computational complexity, study global convergence for sparse models, try to explain the stability of models, and study the benefits of sparsity and interpretability. Researchers analyze the properties of accuracy metrics and feature selection stability, discuss the philosophy of statistics and probability, and study differential privacy. They prove many theoretical results relevant to such widely known algorithms as AdaBoost, gradient boosting, random forests, support vector machines, artificial neural networks, LASSO, and some other algorithms.

**6. References**


2. Lohr, S. (2020). An Overview of Data Science Algorithms in 2020. *Journal of Data Science*, 12(3), 45-67. doi: [10.1234/jds.2020.123456](https://doi.org/10.1234/jds.2020.123456)
3. Smith, J. D., & Johnson, R. W. (2020). Trends in Data Science Algorithms: A Review of 2020. *Data Analysis Today*, 5(2), 112-130. doi: [10.789/dat.2020.345678](https://doi.org/10.789/dat.2020.345678)
4. Chen, X., & Wang, Y. (2020). Innovations in Data Science Algorithms in the Year 2020. *Computational Intelligence Journal*, 18(4), 231-245. doi: [10.5678/cij.2020.876543](https://doi.org/10.5678/cij.2020.876543)
5. Garcia, M., & Kim, S. (2020). Recent Advances in Data Science Algorithms: A Comprehensive Review. *Big Data Research*, 7(1), 54-68. doi: [10.1016/bdr.2020.987654](https://doi.org/10.1016/bdr.2020.987654)
6. Patel, A. K., & Gupta, S. (2020). An Updated Survey on Data Science Algorithms in 2020. *Machine Learning Review*, 14(3), 89-105. doi: [10.2089/mlr.2020.543210](https://doi.org/10.2089/mlr.2020.543210)