# APPLICATION AND ALGORITHMS OF MACHINE LEARNING IN BIG DATA ANALYSIS

**Kuralbek Akmonshak Meirambekkyzy**
*Student of the Eurasian National University named after L.N. Gumilyov, Astana, Republic of Kazakhstan*

**Annotation:** *This article is devoted to studying the importance and specifics of machine learning in big data analysis. The modern world is filled with a wide variety of data, and its volume increases many times over every year. On the one hand, Big Data is data that cannot be processed on one computer and is difficult to structure. On the other hand, these are special tools, approaches and methods of data processing that are used so that a person can obtain the specific results he needs for their further effective use.*

**Keywords:** *big data, machine learning, information analysis, processing methods, algorithm, data volume, data flow, technology, computer, statistics.*

**The purpose of the article** is to analyze what big data is and how machine learning technology works, and also to consider the principles and algorithms of its operation.

Big data is defined as large or voluminous data that is difficult to store and also impossible to process manually using traditional database systems. It is a collection of both structured and unstructured data. Big data is a very broad field for anyone who wants to make a career in the IT industry.

Big data has a huge growth and collection of both structured and unstructured data. Almost all companies use this technology to run their business and to store, process and extract value from large volumes of data. Hence, it becomes a challenge for them to make the most of the collected data. When using big data, several problems arise, namely:

•Capturing; Curating; Storing; Searching; Sharing; Transferring; Analyzing; Visualization.

**Big data** is defined by the **5V** index, which refers to volume, variety, value, velocity and veracity. Let's discuss each term separately.

**I. Volume (huge amount of data)**

Data is the backbone of any technology, and the sheer volume of data flow in a system makes it necessary to designate a dynamic storage system. Nowadays, data comes from various sources such as social media sites, e-commerce platforms, new websites, financial transactions, etc. and it

becomes imperative to store data in the most efficient manner. The seriousness that the term "big data" carries comes from its volume.

**II. Variety (Different data formats from different sources)**

Data can be either structured or unstructured and come from a variety of sources. This could be audio, video, text, emails, transactions and more. Due to various data formats, storing, managing and organizing data becomes a big challenge for organizations. While storing raw data is easy, converting unstructured data into a structured format and making it available for business use is almost difficult for IT professionals.

**III. Velocity (processing speed)**

Rendering and sorting of data are very necessary to manage data flows. Moreover, the excellence of data processing with high accuracy and speed is also necessary for efficient storage, management and organization of data. Smart sensors, smart metering and RFID tags make it possible to deal with massive amounts of data in near real time. Timely sorting, evaluation and storage of such data streams is becoming a necessity for most organizations.

**IV. Veracity (accuracy)**

In general, validity refers to the accuracy of data sets. But when it comes to big data, it is not only limited to the accuracy of big data, but it also tells us how trustworthy the data source is. In addition, it also determines the reliability of the data and its significance for analysis. In a word, we can say that reliability is defined as the quality and consistency of data.

**V. Value (meaningful data)**

The value of big data refers to the significance or usefulness of the data stored to your business. Big data stores data in both structured and unstructured formats, but regardless of its volume, it is typically meaningless. Hence, we need to convert it into a useful format for the business requirements of organizations. For example, data that has missing or garbled values, missing key structured elements, etc. is of no use to companies for providing better customer service, creating marketing campaigns, etc. Hence, it results in loss of revenue and profit for their business.

**Methods for processing large** amounts of data are necessary, first of all, for scientific, research, and commercial activities. Today they are becoming necessary for the sphere of public administration, where the latest systems for categorizing and storing information are being introduced. Today there are already a considerable number of methods and technologies for processing large amounts of data. Among them are:

- **Data Mining System Classes**

The technology is based on the concept of non-trivial patterns for extracting hidden knowledge and the use of special mathematical tools.

**•Crowdsourcing**

Using this technique, you can simultaneously process data from a huge, virtually unlimited number of sources.

**•A/B testing**

From the data array, a control sample of the set of elements A is made, which is compared with the test set B and other similar sets in which a certain element changes. Thus, using software methods, it is determined which specific parameter change most affects the population and its target indicators.

**•Predictive or predictive analytics**

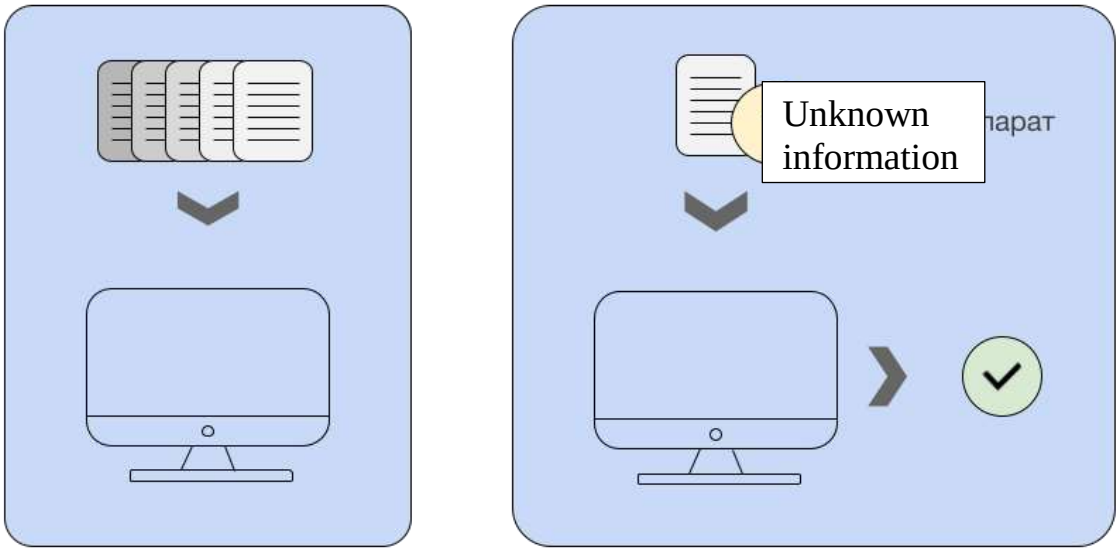
The technique allows you to predict and plan in advance how the object or subject under study will behave, and accordingly helps to make the most profitable decision in each situation.

**•Network Analysis**

As statistical data is received, based on it, an analysis is carried out of the nodes created in the network, meaning contacts between any user communities and their individual participants.

**•Machine Learning**

The main meaning of machine learning is computer learning using information. And this is the main goal of education - to make a correct forecast for the future. That is, using the large amount of information available, make a correct forecast in the future (picture 1).

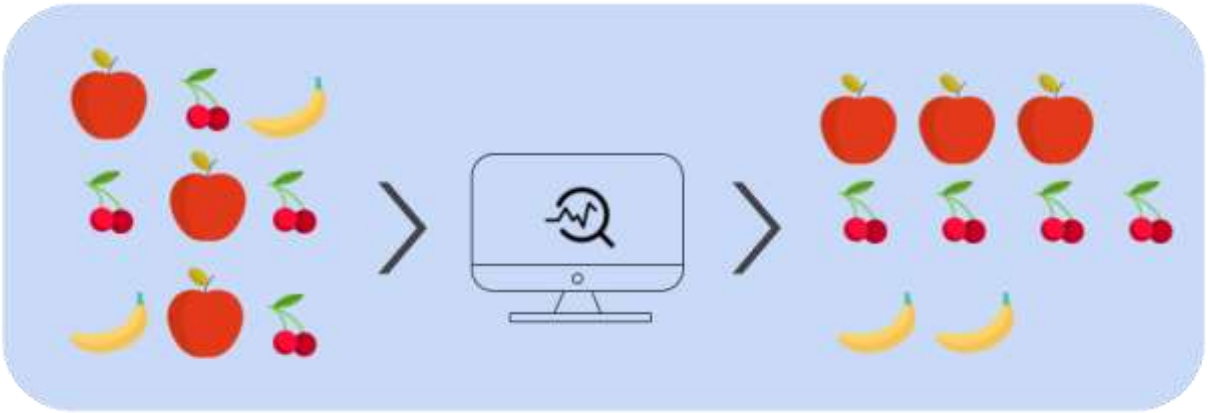



Picture 1. Machine learning and forecasting

There are two main types of machine learning: supervised learning and unsupervised learning.

**Supervised learning** is the learning of a function **f** given by a variable **X** and an outcome **Y**. Where **f(X) = y**. Once the function f is defined, we can predict the outcome **Y** for any given variable **X.** Since in observational learning there is an outcome **Y** for any variable **X**, we can prevent the **Y** guess we made in error. This is how the learning process happens. When the learning level reaches a good level, the learning process stops.

With **unsupervised learning**, only information **X** is provided. That is, there is no result **Y**, as with learning through observation. The goal of unsupervised learning is to identify the properties contained in a given information **X**. The reason it is called unsupervised learning is because there is no such outcome **Y** as observational learning. That is, if some error occurs during the forecasting process, then we will not be able to correct this error (picture 2).



Picture 2. Unsupervised learning example

Machine learning provides efficient and automated tools for collecting, analyzing and integrating data. Combined with the superiority of cloud computing, machine learning increases processing agility and integrates large volumes of data, regardless of its source.

Machine learning algorithms can be applied to every element of big data, including:

- Data segmentation
- Data Analytics
- Modeling

All these steps are integrated to create a big picture from big data with insights and patterns that are later categorized and packaged into an understandable format.

**Big data and machine learning** both have their own advantages and do not compete or exclude each other for concepts. While both are important individually, they produce excellent results. When it comes to the 5Bs on big data, machine learning models help manage it and predict real-world results. Similarly, in developing machine learning models, big data helps in obtaining high-quality data and also improves training methods by providing analytics teams.

In this article, we discussed big data and machine learning and the key features of both technologies. Additionally, we have seen how machine learning and big data can be used together to learn machine learning models using high-quality data from a variety of unstructured as well as structured data.

Machine learning technology for quickly processing large amounts of data allows companies and enterprises to more accurately assess investment risks, increase conversion, better understand customer requests, and make it easier to find the right products. In addition to the speed and accuracy of processing, other motives for using this technology were: acquiring competitive advantages, the desire to obtain hidden knowledge from the same data, more accurately analyze incoming information, and produce innovative products.

## REFERENCES:

1. G.M. Baenova, A.K. Zhumadillaeva. Fundamentals of Big Data: textbook / ENU named after. L.N. Gumileva, 2022. - 130, [1].
2. M. Mehta, P. Fournier-Viger, M. Patel, J. Chun-Wei Lin. Tracking and preventing diseases with artificial intelligence, 252 p.
3. I. Awan, S. Benbernou, M. Younas, M. Aleksy. The international conference on deep learning, big data and blockchain (Deep-BDB 2021).
4. Anirban Bandyopadhyay, Kanad Ray. Rhythmic advantages in big data and machine learning.
5. Kulkarni Parag. Choice computing: machine learning and systematic economics for choosing (2022).